Optimal Frame Duration for Oracle Audio Signal Separation is Determined by Joint Minimization of Two Antagonistic Artifacts

Stephen Voran
Institute for Telecommunication Sciences
Boulder, Colorado, USA
svoran@ntia.gov

***Abstract*—We demonstrate that the optimal audio signal processing frame duration in oracle binary masking and oracle magnitude restoration is determined by joint minimization of two antagonistic artifacts: temporal blurring (which increases with frame duration and log-spectral-error change per unit time (which decreases with frame duration). This is novel — the factors underlying the empirical optimization of frame duration have not been previously identified. Signal stationarity alone cannot explain the existence of an optimal frame duration. Stationarity can explain why a frame duration is too long, but it cannot explain why a frame duration is too short.**

We introduce a method for measuring the stationarity of an audio signal. We then use this essential tool along with measurements, modeling, and analysis in order to identify the two underlying factors that cause there to be an optimal frame duration. In addition we show that when recovering s from the mixture $y = s + n$ with oracle binary masks or oracle magnitudes, the stationarity of s and the stationarity of n have opposite influences on the optimal frame duration. Increasing the stationarity of s increases optimal frame duration but increasing the stationarity of n *decreases* optimal frame duration. Stationarity alone cannot explain these opposing influences but our results do.

***Index Terms*—frame size, oracle binary mask, source separation, speech enhancement, stationarity**

I. Introduction

Separating mixtures of acoustic signals into a desired portion (often speech) and an undesired portion (often environmental sounds or competing speech sources) is an important audio signal processing problem. The problem is commonly addressed by first transforming groups of time-domain samples (called frames) into the frequency domain for further processing. The appropriate time-duration for these frames is selected empirically — frames that are too long or too short produce bad-sounding results. Published observations on optimal frame duration connect it to signal stationarity. Stationarity is indeed a factor, but it fails to fully explain the existence of optimal frame durations.

We show that the optimal audio signal processing frame duration in oracle binary masking and oracle magnitude restoration is determined by joint minimization of two antagonistic artifacts: temporal blurring (which increases with frame duration and log-spectral-error change per unit time (which decreases with frame duration). The use of oracle masks and magnitudes enables us to uncover the intrinsic effects of frame duration, unencumbered by its effects (including frequency resolution) on mask estimation or magnitude estimation. These intrinsic effects are common to both binary mask-based and magnitude-based separation, and are present even in the best-case scenario where binary masks or magnitudes are perfectly estimated. The use of oracles here is properly motivated and it achieves the intended goal, but it may also weaken the connection between these results and practical applications.

We continue with background on the separation problem and discussion of existing results on optimal frame sizes and stationarity of speech signals. Next we discuss the issues inherent in measuring stationarity. We develop the stationarity index ψ which is an essential tool for this work. We present speech quality versus frame duration (τ) curves to demonstrate that the optimum frame duration (τ_{opt}) is an increasing function of signal stationarity and a *decreasing* function of noise stationarity when using an oracle binary mask (OBM) or oracle magnitude recovery (OMR). We report that power concentration, signal lost, and noise admitted respond weakly to τ and cannot account for a significant portion of the measured quality range. Our mathematical analysis of OBM and OMR leads us to a convolutional viewpoint and we find that a very simple convolutional noise model reproduces nearly all of the observed speech quality effects. We then identify, describe, and measure two antagonistic effects. One effect produces worse artifacts with increasing τ and the other effect produces worse artifacts with decreasing τ. Thus quality is maximized by finding the compromise τ value that minimizes the perception of these two artifacts combined.

This paper does not propose a new method for selecting frame durations, it does not suggest new frame durations, and it does not propose new algorithms for separation. Instead it explains the existence of optimal frame durations when the issues of mask or magnitude estimation have been removed from the problem.

II. Background

The earliest work on separating audio signals is [1]–[3], a very small but broad sampling of subsequent contributions can be found in [4]–[9], and the more recent comprehensive

This paper was presented at, and will appear in the proceedings of, the IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP 2021), Tampere, Finland, October 6-8, 2021. MMSP Proceedings are © IEEE.

treatments [10], [11] provide hundreds of additional citations. Machine learning approaches are effective (e.g., [12]–[14]) and may eventually eliminate the need to understand signal processing details but there remains value in better understanding the properties of the classic building blocks of the field, as in this paper.

Time-frequency representations are key to this work. Let $\boldsymbol{x}$ be a vector of real-valued time-domain audio samples and $\boldsymbol{X}$ be the corresponding matrix of complex time-frequency samples. We use

$$\boldsymbol{X} = F(\boldsymbol{x}, \tau, \tau_S, \tau_T, f_s), \quad \boldsymbol{x} = F^{-1}(\boldsymbol{X}, \tau, \tau_S, \tau_T, f_s) \quad (1)$$

to concisely represent this relationship, parametrized by frame duration τ, frame stride τ_S, and DFT length τ_T (all in units of time) and sample rate f_s. This parameterization emphasizes that we are concerned with the *time duration* of frames rather than the *number of samples* in frames. F applies a square root Hann window after framing and before DFT. F^{-1} applies the inverse DFT followed by the square root Hann window as part of the overlap-and-add (OLA) process. This allows exact reconstruction of $\boldsymbol{x}$ with the exception of a fraction of the initial and final frames. Each column of $\boldsymbol{X}$ contains $\frac{f_s}{2} \cdot \tau_T + 1$ complex frequency-domain samples that cover DC to Nyquist.

Given signal $\boldsymbol{s}$, noise $\boldsymbol{n}$, and their sum $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n}$, use (1) to produce $\boldsymbol{S}$, $\boldsymbol{N}$, and $\boldsymbol{Y}$. In mask-based separation a binary mask is multiplied (element-by-element) with $\boldsymbol{Y}$ to form $\hat{\boldsymbol{S}}$, thus passing some complex values in $\boldsymbol{Y}$ unchanged and replacing others with zero. The challenge is to estimate a mask that makes $\hat{\boldsymbol{S}}$ a good approximation to $\boldsymbol{S}$. Alternately, when magnitude recovery is used, we modify the magnitudes of the values in $\boldsymbol{Y}$ to produce $\hat{\boldsymbol{S}}$ with the goal that they approximate the magnitudes in $\boldsymbol{S}$. The phases in $\hat{\boldsymbol{S}}$ remain the same as those in $\boldsymbol{Y}$. In this approach the challenge is to estimate the magnitudes in $\boldsymbol{S}$ so they can be used in $\hat{\boldsymbol{S}}$. In either approach the output is $\hat{s} = F^{-1}(\hat{\boldsymbol{S}}, \tau, \tau_S, \tau_T, f_s)$.

The literature supporting these two families of work is vast and researchers have empirically found the range of frame durations τ that provide the best sounding results, while also considering algorithmic delay and efficient (power-of-two) transform lengths. Discussion in [11] points to Fig. 5 in [15] which shows that maximum signal-to-distortion ratio is achieved for $\tau = 54$ ms (1200 smp) for speech and 186 ms (4100 smp) for music in the case of separation by oracle binary masks in the modified discrete cosine transform domain (f_s = 22,050 smp/s). The authors of [15] offer that "This is likely to be because music is more 'stationary' than speech."

In [16] three figures of merit were considered in the mask-based separation of speech signals with f_s = 16,000 smp/s. All three were maximized by frame size 1024 smp, corresponding to $\tau = 64$ ms. Three other figures of merit were applied to a two-channel speech separation algorithm ($f_s = 11{,}025$ smp/s) in [17]. A search for optimal frequency resolution led to the range 10 to 20 Hz, corresponding to the range $\tau = 46$ to 93 ms. The authors add that "The average stationary period of a speech signal is around 40 ms ..." and they refer to [18].

Together these empirical optimization results for speech suggest frame durations between 46 and 93 ms. Values in this range are commonly seen in published algorithms and authors commonly suggest connections between these results and the stationarity of the signals.

Some numerical statements on the stationarity of speech have been published as well. A statement regarding the applicability of the "DTFT representation" in speech processing [10] (p. 31) gives the range 10 to 30 ms because during that time interval "...the properties of speech do not change much." Reference [18] mentions several values between 15 and 40 ms as suitable frame durations for speech processing. In the specific context of autocorrelation and Fourier transforms [18] states that (p. 54) "...20 and 40 ms often bring good results for female and male voices, respectively." If we interpret these as statements on the stationarity of speech, the full range of stationarity values reported covers 10 to 40 ms.

The intuition that the optimal frame duration τ_{opt} is linked to signal stationarity is sound but incomplete. The relationship is far from trivial and the ways in which signal stationarity drive τ_{opt} is a topic worthy of investigation. First note that the ranges reported for stationarity (10 to 40 ms) and τ_{opt} (46 to 93 ms) are actually *disjoint*. Note also that while stationarity may provide an argument for an upper limit on suitable τ values, it does not speak to a lower limit.

Additional proof that the relationship between stationarity and τ_{opt} needs explanation is provided in the following example. Given $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n}$, the task of recovering $\boldsymbol{s}$ from $\boldsymbol{y}$, and the ability to use oracle magnitudes or masks, independently adjust the stationarity of $\boldsymbol{s}$ and $\boldsymbol{n}$ and find the optimal frame duration τ_{opt}. As the stationarity of $\boldsymbol{s}$ increases τ_{opt} increases as well. But as the stationarity of $\boldsymbol{n}$ increases τ_{opt} *decreases*. We will explain the sources of these opposing trends by identifying the additional, more immediate, factors that actually drive τ_{opt}.

III. Measuring Stationarity

A stochastic process $x(t)$ is stationary if "$x(t)$ and $x(t + c)$ have the same statistics for any c" [19]. Any signal produced by a physical process is always evolving at least minutely and measurement noise is always present. This means that the statistics of a real audio signal can never be exactly "the same." But we can adopt thresholds in order to identify temporal regions where audio signal statistics are "approximately the same" and call these regions of "approximate stationarity."

Audio signal statistics show slow evolution across some time windows and abrupt changes at other points in time. So in any audio signal the lengths of the regions of approximate stationarity will cover a distribution. When a single value of "stationarity" is needed it is natural to convert this distribution to a value by using a measure of central tendency. Thus when a single time value describes the stationarity of an audio signal, that value would be most precisely labeled as a "mean approximate stationarity value," for example. Such values will depend on the specific statistics used, the number of audio samples used to compute those statistics, the thresholds adopted, and the measure of central tendency used.

We developed a simple yet effective stationarity index ψ. We provide a narrative description here and we provide pseudo code at [20]. We based ψ on log variance and the lag 1 to 12 autocorrelation coefficients since they are very descriptive, relevant, and ubiquitous in audio signal processing. We calculate each statistic over 5 ms. This window length represents a compromise between robust statistics (longer windows) and temporal resolution (shorter windows). We advance that window one sample for each new calculation. The result is thirteen sequences of statistics. Next we find contiguous blocks of windows where none of the thirteen statistics changes by more than the specified thresholds. We then filter the resulting blocks to remove very low power blocks because these are irrelevant from an audio signal perspective. We also remove blocks that are shorter than 10 ms because in these blocks the 5 ms windows are overlapping and that overlap compromises the tests for changes in statistics between windows.

The fraction (R_b) of the original audio samples that are in a surviving block is a unitless measure of the level of approximate stationarity. The average duration of the surviving blocks (L_b) is a measure of the average duration of approximate stationarity. It is natural and useful to combine the level and duration factors to arrive at an index of stationarity $\psi = R_b \cdot L_b$ which has units of time.

The index ψ is an essential tool for the work that follows. It is objective and quantitative and it produces values consistent with those less objective values seen in the literature surrounding statements about "typical" or " average" speech stationarity, or speech "not changing much" over a time interval. We computed ψ for each of 260 recordings from 13 different talkers, each saying 20 sentences taken from [21]. Values ranged from 7 to 29 ms and per-talker averages ranged from 10 to 20 ms. Two male talkers recorded the same 20 sentences and consistent with intuition, the more relaxed talker produced an average ψ value of 20 ms while the talker who rushed produced an average ψ value of 13 ms. We created lower and higher stationarity speech sets by sorting the 260 recordings according to increasing ψ. We defined the first and last quarters (65 recordings each) to be the sets Ψ_L (ψ range 7 to 12 ms, mean 9.7 ms) and Ψ_H (ψ range 19 to 30 ms, mean 21.8 ms).

We analyzed noise recordings and found ψ values of 13 ms for high-activity office noise, 24 ms for coffee shop noise, and 61 ms for monotonous sounding saw noise. This trend agrees with perception. As expected, ψ goes to ∞ (or length of signal) in the case of computer-generated white noise. Adding noise to speech can increase or decrease ψ depending on the noise type. Analysis of music examples also produced intuitive results. We found ψ values of 13 ms for up-tempo electronic percussive dance music, 18 ms for percussive jazz, 60 ms for a slow waltz played by strings, and 188 ms for a sustained modulated flute note.

IV. Quality as a function of Frame Duration

We seek to understand what limitations the choice of frame duration τ places on the binary mask and magnitude recovery algorithms. Toward that end we use oracle masks or oracle magnitudes. There are countless techniques for estimating masks or magnitudes and each may display its own response to τ. By adopting oracles we eliminate any assumptions about the behavior of these estimation techniques. Mask and magnitude estimation certainly depend on τ, but we show that a very strong τ dependence remains after they have been completely removed from the problem.

Consider $\boldsymbol{s}$, $\boldsymbol{n}$, and $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n}$, along with $\boldsymbol{S}$, $\boldsymbol{N}$, and $\boldsymbol{Y}$ defined via (1). The 0 dB OBM $\boldsymbol{M}$ has the same size as $\boldsymbol{Y}$ and each element is defined as

$$\begin{aligned} |\boldsymbol{S}_{ik}| > |\boldsymbol{N}_{ik}| \implies \boldsymbol{M}_{ik} &= 1, \\ \text{otherwise,} \qquad \boldsymbol{M}_{ik} &= 0. \end{aligned} \tag{2}$$

Element-by-element multiplication applies the mask $\boldsymbol{M}$ to $\boldsymbol{Y}$ to produce $\hat{\boldsymbol{S}}$. The mask passes only those time-frequency elements of $\boldsymbol{Y}$ where the power in $\boldsymbol{S}$ exceeds the power in $\boldsymbol{N}$. The time-domain OBM output is produced by

$$\hat{\boldsymbol{s}}_{OBM} = F^{-1}(\boldsymbol{Y} \cdot \boldsymbol{M}, \tau, \tau_S, \tau_T, f_s). \tag{3}$$

In OMR, $\hat{\boldsymbol{S}}$ is formed from the magnitude of $\boldsymbol{S}$ and the phase of $\boldsymbol{Y}$,

$$\hat{\boldsymbol{s}}_{OMR} = F^{-1}(|\boldsymbol{S}| \cdot e^{j\angle \boldsymbol{Y}}, \tau, \tau_S, \tau_T, f_s), \tag{4}$$

where phase angle extraction ($\angle$), multiplication, and exponentiation are again element-by-element.

To study the effect of τ we used the 65 lower-stationarity speech signals in Ψ_L (mean $\psi = 9.7$ ms) and 65 higher-stationarity speech signals in Ψ_H (mean $\psi = 21.8$ ms). We mixed speech with coffee shop noise ($\psi = 24$ ms), saw noise ($\psi = 61$ ms), and white noise ($\psi \to \infty$ ms) at 0 dB SNR.

We applied OBM and OMR and evaluated the resulting speech quality using the POLQA [22] and the wideband PESQ [23] speech quality estimation algorithms. PESQ results showed the same trends as POLQA results, so for conciseness and clarity we display only POLQA results here. Thus "optimal" refers to a peak in POLQA or PESQ speech quality values, confirmed by informal listening tests. We repeated all work at $f_s =$ 48,000, 16,000, and 8000 smp/s to confirm that frame duration, not samples-per-frame, was driving results. Peaks appear at similar τ values and small differences are consistent with the bandwidth limitations imposed by the lower sample rates. For conciseness and clarity we show only results for 48,000 smp/s.

To mitigate time-domain aliasing [24] we experimented with DFT lengths τ_T= τ, 2τ, and 4τ (accomplished by zero padding). We found that moving from τ to 2τ increased quality measurably, but increasing to 4τ showed no clear additional benefit. Thus we adopted τ_T= 2τ. We also experimented with frame stride $\tau_S = \tau/2, \tau/4$, and $\tau/8$. We found that stride influences speech quality much more weakly than transform length and we adopted the commonly used value $\tau_S = \tau/2$.

Fig. 1 shows speech quality values for six cases of speech and noise in the OMR case. The nominal scale for the POLQA MOS-LQO (speech quality) values is 1 (bad) to

5 (excellent), so peaks can be used to find τ_{opt}. For all three noise environments the more stationary speech in Ψ_H produces slightly higher values of τ_{opt} than the less stationary speech in Ψ_L. Coffee shop noise produces the highest values of τ_{opt} and saw noise produces slightly *lower* values, due to its *higher* stationarity. White noise produces *even lower* values of τ_{opt}, due to its *much higher* stationarity. The quality results can be confirmed by listening to example audio files [20]. The OBM case produced similar results.

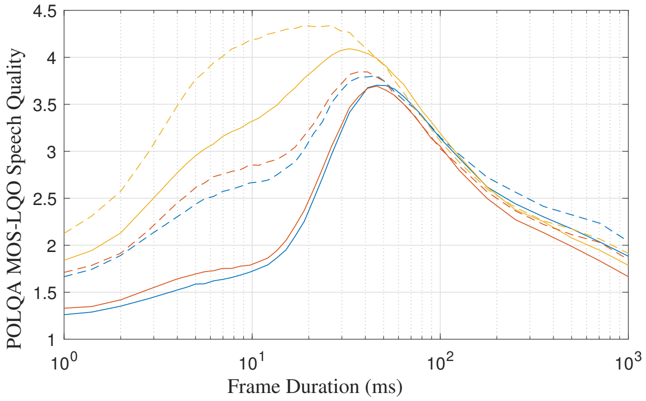

Fig. 1. Speech quality as a function of frame duration (τ) for OMR case. Noise types are coffee shop (blue), saw (red), and white (gold). Dashed and solid lines show lower and higher stationarity speech (Ψ_L and Ψ_H) respectively. See text discussion on opposing trends in τ_{opt} with respect to stationarity of signal and noise.

We also measured signal power removed by the mask, noise power admitted by the mask, and (inspired by [17]) power concentration, each as a function of τ. These show weak effects with extrema located in similar but not identical locations as in Fig. 1. Signal power removed, noise power admitted, and power concentration measure how amenable signals are to separation in the oracle case, and there is indeed a mild dependency on τ. However the range of quality shown in Fig. 1 is dramatic and not consistent with the modest changes in signal lost, noise admitted, or power concentration. In a separate test we adapted mask thresholds and input SNR so that signal lost and noise admitted remained constant as we varied τ. Resulting quality curves showed nearly all of the quality range seen in Fig. 1. These results make it clear that τ drives artifact levels directly in a strong way, and has a much weaker indirect influence via signal separability.

V. Simple Model Captures Majority of Effects

To understand the main factors that drive τ_{opt}, we simplify the situation by invoking a model that closely reproduces the results seen in Fig. 1. OBM and OMR can be unified in the sense that in both cases the time-frequency representations of the original signal $\boldsymbol{S}$ and the recovered signal $\hat{\boldsymbol{S}}$ are related through element-by-element multiplication

$$\hat{\boldsymbol{S}} = \boldsymbol{S} \cdot \boldsymbol{G}. \tag{5}$$

In the OBM case

$$\boldsymbol{G_{ik}} = (1 + \boldsymbol{N}_{ik}/\boldsymbol{S}_{ik}) \cdot \boldsymbol{M}_{ik}, \tag{6}$$

and in the OMR case

$$\begin{gathered}\boldsymbol{G_{ik}} = e^{j\boldsymbol{\phi}_{ik}},\ \boldsymbol{\phi}_{ik} = \arctan(\sin(\theta_{ik}), \frac{|\boldsymbol{S}_{ik}|}{|\boldsymbol{N}_{ik}|} + \cos(\theta_{ik})), \\ \theta_{ik} = \angle\boldsymbol{N}_{ik} - \angle\boldsymbol{S}_{ik},\end{gathered} \tag{7}$$

where $\arctan(\ \cdot\ ,\ \cdot\ )$ indicates the four-quadrant arctangent function. The elements of $\boldsymbol{G}$ are complex. In the OBM case they satisfy $0 \leq |\boldsymbol{G}_{ik}| \leq 2$. In the OMR case they satisfy $|\boldsymbol{G}_{ik}| = 1$.

We can apply the inverse DFT to each column of each matrix in (5) to produce matrices $\hat{\boldsymbol{S}}_t$, $\boldsymbol{S}_t$, and $\boldsymbol{G}_t$. Each of these contains a single time-domain frame in each column and OLA can then convert these to time-domain vectors. Frequency-domain multiplication is equivalent to circular convolution in the time domain, so (5) can be written

$$\hat{\boldsymbol{S}}_t = \boldsymbol{S}_t * \boldsymbol{G}_t, \tag{8}$$

where $*$ indicates circular convolution of the corresponding columns of $\boldsymbol{S}_t$ and $\boldsymbol{G}_t$. Equation (8) makes it clear that each recovered frame is an original frame convolved with a noise source that is the inverse DFT of (6) or (7). As expected, both expressions for $\boldsymbol{G}$ depend on SNR. In both the OBM and OMR cases, as the noise vanishes $\boldsymbol{G}$ goes to all ones and $\boldsymbol{G}_t$ becomes the convolutional identity.

For OMR the power in $\boldsymbol{G}$ (7) is exactly constant, hence the power in $\boldsymbol{G}_t$ is as well. Yet the choice of τ can cause the resulting quality to range from near bad to near excellent. In OBM the power in $\boldsymbol{G}$ (6) is self-limiting. As input noise increases, $|\boldsymbol{N}_{ik}|$ / $|\boldsymbol{S}_{ik}|$ tends to increase but the mask $\boldsymbol{M}$ contains more zeros and this compensates. We reduced the input SNR from +40 to -10 dB (a 50 dB increase in noise) and the average power in $\boldsymbol{G}_t$ increased by just 19 dB (with $\tau =$ 21.3 ms). The average power in $\boldsymbol{G}_t$ is also largely invariant to τ. Using coffee shop noise at 0 dB SNR, we swept τ from 1 to 1000 ms and the average power in $\boldsymbol{G}_t$ changed by just 2.8 dB. And for $10 \leq \tau \leq 100$ ms the power changes by only 0.9 dB. We have experimented with holding the power in $\boldsymbol{G}_t$ constant while changing τ, and speech quality results are similar to those in Fig. 1. This shows that changes in the power of the noise $\boldsymbol{G}_t$ do not explain the existence of τ_{opt}.

So for both the OBM and OMR cases it is clear that noise power is not driving quality. Therefore quality must be driven by either the distribution of noise values in $\boldsymbol{G}_t$ or by the frame duration τ. In order to separate these two factors we next introduce a model for the convolutional noise $\boldsymbol{G}_t$ that uses distributions that are independent of τ.

We analyzed $\boldsymbol{G}_t$ for both OBM and OMR in the case of coffee shop noise at 0 dB SNR. The distributions are similar and are easy to model without any dependence on τ. In this model, all samples are normal, the first sample of each frame is drawn from $\mathcal{N}(0.5, 0.3)$, and all other samples are drawn from $\mathcal{N}(0.0, 0.005)$. Each frame is scaled so the sum of the squared noise samples is one. (Note that if the first sample were one and all other samples were zero, $\boldsymbol{G}_t$ would be the convolutional identity.)

We used this model to produce noise $\hat{\boldsymbol{G}}_t$ and then used circular convolution to corrupt clean speech $\boldsymbol{S}_t$ with this modeled convolutional noise analogous to (8). Finally, we used OLA to produce a time-domain speech signal:

$$\hat{\boldsymbol{s}}_{Model} = \text{OLA}(\boldsymbol{S}_t * \hat{\boldsymbol{G}}_t). \tag{9}$$

Thus $\hat{\boldsymbol{s}}_{Model}$ has no τ dependence other than the fundamental and unavoidable dependence. That is, τ cannot be influencing power concentration, separability, or frequency resolution because all of these have been eliminated in the calculation of $\hat{\boldsymbol{s}}_{Model}$. In addition the power and distribution of the convolutional noise $\hat{\boldsymbol{G}}_t$ are independent of τ. The *only* way that $\hat{\boldsymbol{s}}_{Model}$ depends on τ is through the duration of the circular convolutions.

Fig. 2 shows that this model reproduces the τ sensitivity measured in the OBM and OMR cases. This highly-simplified model reproduces the peaks near 46 ms and even much of the detail that differentiates the groups Ψ_L and Ψ_H. The artifacts produced sound similar to the actual artifacts and this can be confirmed by listening to files we have provided [20] and comparing the sounds.

Fig. 2 and the demonstration [20] show that the simplified situation of (9) is clearly sufficient to represent the most important factors driving τ_{opt} in the OBM and OMR cases. The only τ dependence in (9) comes from an innate and fundamental property — the duration of the circular convolutions. We have eliminated all other influences of τ and we have shown that frame duration in the purest sense is a strong driver of speech quality. We next characterize this fundamental driving force in terms of two opposing artifacts.

VI. Antagonistic Artifacts Determine Optimal Frame Duration

The audio examples [20] for (9) reveal two antagonistic artifacts with very distinct sounds. One worsens when τ is increased and the other worsens when τ is decreased. Increasing τ increases temporal blurring. This is different from time-domain aliasing which has been made imperceptible by adopting $\tau_T = 2\tau$. Circular convolution adds time-shifted scaled copies of the signal and this temporal blurring over the duration of the frame produces reverberation-like or "phasy" artifacts. A more stationary signal exhibits less change within a frame and this blurring will be less audible [20]. So the temporal blurring artifact (which drives τ_{opt} downward) is weakened with increased stationarity of the signal. Thus increased signal stationarity allows τ_{opt} to increase.

To quantify this artifact we developed envelope infill $\xi(\tau)$ which measures power levels in areas where the clean speech was silent. We compute speech envelopes for $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$ similar to in [25]. Where the envelope of $\boldsymbol{s}$ falls at least 30 dB below the peak value we measure the power in $\hat{\boldsymbol{s}}$. Averaging this power produces $\xi(\tau)$ which tells the extent to which these quiet areas have been "filled in."

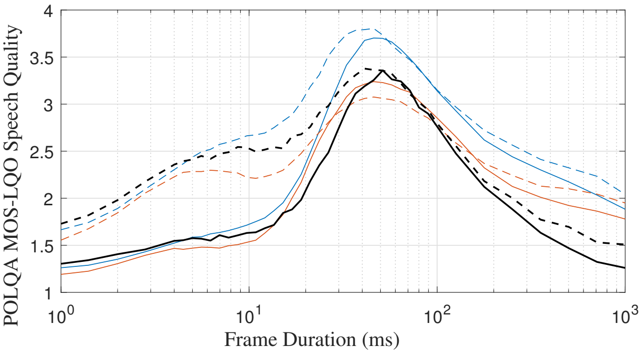

Fig. 2. Speech quality as a function of frame duration (τ) for 0 dB SNR coffee shop noise. OBM (3), OMR (4), and model (9) shown in blue, red, and black, respectively. Dashed and solid lines show lower and higher stationarity speech (Ψ_L and Ψ_H) respectively. Model captures vast majority of OBM and OMR quality effects.

We now turn to the effects of decreasing τ. This causes increasing amounts of a harsh granular (signal-correlated) noise that sounds somewhat like the artifact produced by the MNRU [26]. We can also simulate this sound and its τ dependence by multiplying randomly selected time-domain frames (columns of $\boldsymbol{S}_t$) by -1. This simulation is easily understood — a constant phase inversion is not an audible impairment but *changing* phase between normal and inverted is audible and the more frequently this occurs, the more audible and annoying it becomes. We also know that changes in spectral shaping are more noticeable than fixed spectral shaping and thus are key drivers of speech quality [27]–[31]. More frequent changes and larger changes will be more audible and annoying.

Taken together, these observations suggest the artifacts caused by decreasing τ are due to unnatural temporal changes in spectra. This motivates us to measure spectral changes across time. Specifically, we measure the log-spectral-error (LSE) [32] between $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$ then average the magnitude of the temporal changes in this LSE to produce $\Delta_t|\text{LSE}(\tau)|$. This is a measure of unnatural and potentially annoying temporal variation in signal spectra.

A more stationary noise $\boldsymbol{n}$ causes less temporal fluctuation in the mask (OBM case) or the phases of $\hat{\boldsymbol{S}}$ (OMR case). This translates to less unnatural temporal variation in the spectrum of $\hat{\boldsymbol{s}}$ and a lower value of $\Delta_t|\text{LSE}(\tau)|$. So the artifact that drives τ_{opt} upward is weakened with increased stationarity of the noise. Thus increased noise stationarity allows τ_{opt} to decrease.

Fig. 3 shows $\xi(\tau)$ and $\Delta_t|\text{LSE}(\tau)|$ for the model (8) used in Fig. 2. These results are averages over the set of 130 speech files formed by merging Ψ_L and Ψ_H. $\xi(\tau)$ is non-decreasing in τ. It is relatively constant below 20 ms and increases above that point. Conversely, $\Delta_t|\text{LSE}(\tau)|$ is non-increasing in τ, dropping most dramatically up to 100 ms, then dropping more slowly above that point. Both $\xi(\tau)$ and $\Delta_t|\text{LSE}(\tau)|$ have units of power in dB.

Fig. 3 also shows that joint minimization of the weighted functional $\xi(\tau) + 0.5 \cdot \Delta_t|\text{LSE}(\tau)|$ produces a minimum near

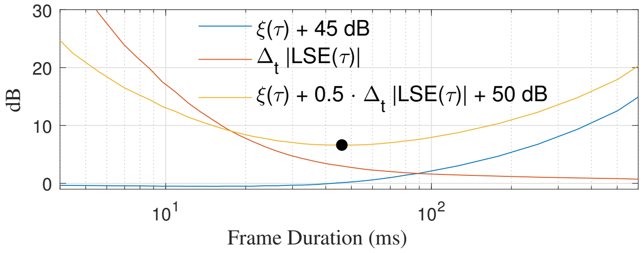

Fig. 3. Antagonistic artifacts measured by $\xi(\tau)$, $\Delta_t|\text{LSE}(\tau)|$, and their weighted sum, shifted for display. Minimum in weighed sum at $\tau = 46$ ms gives maximum speech quality as in Fig. 2.

$\tau = 46$ ms. In other words, if the average perceptual importance of $\Delta_t|\text{LSE}(\tau)|$ is half that of $\xi(\tau)$, then the total artifacts are minimized and the quality is maximized when the frame duration is near 46 ms. This is similar to the location of the speech quality peak associated with (9) in Fig. 2 ($\tau = 40$ ms) and these locations could be matched exactly by placing a slightly smaller weight on $\Delta_t|\text{LSE}(\tau)|$. This result confirms that $\xi(\tau)$ and $\Delta_t|\text{LSE}(\tau)|$ do indeed capture the artifacts that drive optimal frame duration.

VII. Conclusion

Until now, optimal frame durations for separation have never been properly connected to signal stationarity (which can drive upper but not lower limits on frame duration). We have developed a meaningful measure of signal stationarity and have used it to show that speech stationarity and noise stationarity drive optimal frame duration in opposite directions. Our mathematical modeling yields a simple but accurate convolutional model that shows optimal frame durations in the OBM and OMR cases are driven by joint minimization of two competing artifacts — temporal blurring and unnatural temporal variation of spectra. Changes in the levels of stationarity in signal and noise modify the relationship between these two artifacts and thus increase or decrease the optimal frame duration.

References

[1] S. Boll, "Suppression of noise in speech using the SABER method," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, Apr. 1978, pp. 606–609.

[2] ——, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, Apr. 1979, pp. 208–211.

[4] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, Jul. 1998.

[5] L.P. Yang and Q.J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *Journal of the Acoust. Society of America*, vol. 117, Mar. 2005.

[6] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Noise reduction based on adaptive beta-order generalized spectral subtraction for speech enhancement," in *Proc. Interspeech 2007*, Aug. 2007, pp. 802–805.

[7] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1770–1779, Aug. 2011.

[8] T. Virtanen, J. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, Mar. 2015.

[9] S. Voran, "The selection of spectral magnitude exponents for separating two sources is dominated by phase distribution not magnitude distribution," in *Proc. 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2017, pp. 279–283.

[10] P. Loizou, *Speech Enhancement, Theory and Practice*. Boca Raton, Florida: CRC Press, 2013.

[11] T. Virtanen, E. Vincent, and S. Gannot, "Time-frequency processing: Spectral properties," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Hoboken, New Jersey: Wiley, 2018, ch. 2, pp. 15–29.

[12] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[13] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, May 2019.

[14] T. Nakamura, S. Kozuka and H. Saruwatari, "Time-domain audio source separation with neural networks based on multiresolution analysis," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 1687–1701, April 2021.

[15] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933 – 1950, Aug. 2007.

[16] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[17] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

[18] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*. New York: Marcel Dekker, 1989.

[19] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1984.

[20] "Audio demos for frame duration study," www.its.bldrdoc.gov/audio.

[21] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, Sep. 1969.

[22] Recommendation ITU-T P.863 (2018), *Perceptual objective listening quality prediction*, International Telecommunication Union, Geneva, Switzerland.

[23] Recommendation ITU-T P.862.2 (2007), *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, International Telecommunication Union, Geneva, Switzerland.

[24] J. A. Moorer, "A note on the implementation of audio processing by short-term Fourier transform," in *Proc. 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2017, pp. 156–159.

[25] S. Voran, "A bottom-up algorithm for estimating time-varying delays in coded speech," in *Proc. 3rd Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, May 2004.

[26] Recommendation ITU-T P.810 (1996), *Modulated noise reference unit (MNRU)*, International Telecommunication Union, Geneva, Switzerland.

[27] H. Petter Knagenhjelm and W. Bastiaan Kleijn, "Spectral dynamics is more important than spectral distortion," in *1995 Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, May 1995, pp. 732–735.

[28] S. Voran, "Objective estimation of perceived speech quality — Part I: Development of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 4, pp. 371–382, Jul. 1999.

[29] ——, "Objective estimation of perceived speech quality — Part II: Evaluation of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 4, pp. 383–390, Jul. 1999.

[30] S. Voran, "Subjective ratings of instantaneous and gradual transitions from narrowband to wideband active speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Mar. 2010, pp. 4674 –4677.

[31] S. Voran and A. Catellier, "When should a speech coding quality increase be allowed within a talk-spurt?" in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, May 2013, pp. 8149–8153.

[32] S. Voran, "Advances in objective estimation of perceived speech quality," in *Proc. 1999 IEEE Workshop on Speech Coding*, Jun. 1999, pp. 138–140.