Multizone Reproduction of Speech Soundfields: A Perceptually Weighted Approach

Jacob Donley and Christian Ritz
School of Electrical, Computer and Telecommunications Engineering, University of Wollongong
Wollongong, NSW, Australia, 2522
E-mail: jrd089@uowmail.edu.au, critz@uow.edu.au

***Abstract*—In this paper a method for the reproduction of multizone speech soundfields using perceptual weighting criteria is proposed. Psychoacoustic models are used to derive a space-time-frequency weighting function to control leakage of perceptually unimportant energy from the bright zone into the quiet zone. This is combined with a method for regulating the number of basis planewaves used in the reproduction to allow for an efficient implementation using a codebook of predetermined weights based on desired soundfield energy in the zones. The approach is capable of improving the mean squared error for reproduced speech in the bright zone by -10.5 decibels. Results also show that the approach leads to a significant reduction in the spatial error within the bright zone whilst requiring 65% less loudspeaker signal power for the case where the soundfield in this zone is in line with, and hence partially directed to, the quiet zone.**

I. Introduction

Spatial audio reproduction gives listeners a full experience of the acoustic environment, including the sound source, and has been further extended to multizone soundfield reproduction, which provides audio in spatially separated regions from a single set of loudspeakers, originally proposed in [1]. They may also be used for suppressing, or cancelling, audio outside a targeted listening zone [2]. The multizone approach has many applications such as the creation of personal sound zones in multi-participant teleconferencing, entertainment/cinema and vehicle cabins where personal sound zones are optimised to provide one, or many, listener(s) with individual acoustic material [3].

In order to keep the sounds zones personal it is necessary to minimise the interzone interference to maximise the individualistic experience. Some of the earlier methods treat the interference with hard constraints and attempt to completely remove it [1], [4]. This results in zones that are mostly free of the interference, however, this is difficult to achieve in situations where a desired soundfield in the bright zone is obscured by or directed to another zone, as the system requires reproduction signals many times the amplitude of what is reproduced within any zone. This is known as the *occlusion problem* [1], [3], [5] and has been dealt with in various ways such as the control of planarity [6], orthogonal basis planewaves [7] and alleviated zone constraints [7], [8].

Requiring large signals in relation to the reproduced zones means the system is inefficiently directing its energy for the multizone reproduction, with most sound energy present in unattended regions. This may be undesirable at times where listeners commute between sound zones and could put unnecessary strain on loudspeaker drivers. More recent work has focused on alleviating the constraint such that the interference (or leakage) is allowed into other zones, though, the amount can be controlled with a weighting function [7], [8]. Allowing the sound to leak into other zones can improve the practicality of the system but decrease the individuality of zones.

Existing methods focus on single frequency soundfields, although there has been work attempting to create multizone soundfields for wideband speech [9]. More recently, work has been done [10] to extend a method [7] to the reproduction of weighted wideband speech soundfields whilst efficiently maintaining the weighting function in the spatial, time and frequency domain. This allows for dynamic weighting of the zones as well as individual frequency components in time thus allowing each zone's acoustic content to be controlled.

The control of acoustic components to enhance the perception of a signal has been researched thoroughly for applications such as compression [11]. The relationship between the quality in the bright zone and interference in other zones has been subjectively tested [12], however, the occlusion problem is not directly addressed and the planarity control does not consider human perception. Hence, perceptual models are employed in this work in order to enhance the experience in personal sound zones, especially where the occlusion problem is present. Leaked sound energy is treated as unwanted noise in other zones and controlled such that it is perceptually less noticeable as indicated by established psychoacoustic models.

We begin with an explanation of the weighted multizone soundfield method used in this work in Section II. Psychoacoustic models are introduced in Section III as well as the need for regulating the number of weighted basis planewaves. Results of the perceptual weighting and conclusions are given in Section IV and Section V, respectively.

II. Weighted Multizone Wideband Soundfields

The multizone soundfield reproduction layout considered in this work is shown in Fig. 1 and contains a reproduction region, $\mathbb{D}$, which has a radius R. The reproduction region consists of three zones called the bright, quiet and unattended zones which are represented by $\mathbb{D}_b$, $\mathbb{D}_q$ and $\mathbb{D} \cap (\mathbb{D}_b \cup \mathbb{D}_q)'$, respectively. The centres of $\mathbb{D}_b$ and $\mathbb{D}_q$ have a distance of r_z from the centre of $\mathbb{D}$ and each of these zones has a radius of

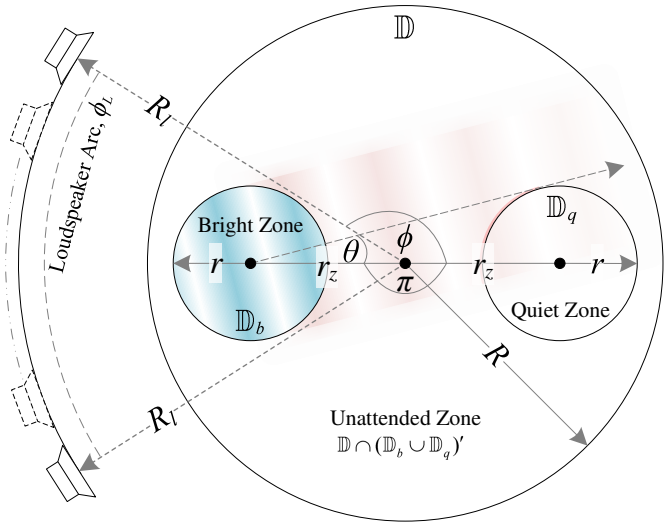

Fig. 1: A weighted multizone soundfield reproduction layout is shown. The figure depicts a situation where the desired soundfield in the bright zone is partially directed towards the quiet zone causing the occlusion problem.

r. Loudspeakers are positioned with a distance of R_l from the centre of $\mathbb{D}$ on an arc of angle ϕ_L which starts at angle ϕ and reproduces planewave speech soundfields in $\mathbb{D}_b$ with an angle of θ.

In the method of weighting multizone soundfields [7], a spatial weighting filter as a function of space, $w(\mathbf{x})$, is used to control the reproduction of sound within each of the zones. Subsequent work [10] extended this approach to allow for space-time-frequency dependent weighting functions, $w(\mathbf{x}, n, k)$, which allows for weighting functions to be adapted based on the signal characteristics of the target soundfield. We denote w_b, w_q and w_u as the weights for $\mathbf{x}_b \in \mathbb{D}_b$, $\mathbf{x}_q \in \mathbb{D}_q$ and $\mathbf{x}_u \in \mathbb{D} \cap (\mathbb{D}_b \cup \mathbb{D}_q)'$, respectively. The reproduced soundfield pressure at any point in the reproduction region is defined as the sum of space-time-frequency dependent weighted soundfield values [10],

$$\hat{p}_w(\mathbf{x}, n) = \sum_{k}^{K} S_w^a(\mathbf{x}, n, k) \tag{1}$$

where $S_w^a(\mathbf{x}, n, k) = \mathrm{f}\left(S^d(\mathbf{x}, n, k), w(\mathbf{x}, n, k)\right)$ is a reproduced soundfield, which is derived as a function of a desired soundfield, $S^d(\mathbf{x}, n, k)$, and a weighting function, $w(\mathbf{x}, n, k)$ using the approaches outlined in [7], [10]. Here, $\mathbf{x}$ is a given position, n is a given time and k is a given frequency. $S_w^a(\mathbf{x}, n, k)$ is summed for K different sinusoidal components. In this work $k = 2\pi f/c$ and $c = 343\,\mathrm{m\,s^{-1}}$.

In (1) $w(\mathbf{x}, n, k)$ allows independently weighting soundfield components in space and time. It is then possible to define the reproduced space-time-frequency domain signal for a particular input as [10],

$$\hat{Y}_w(\mathbf{x}, n, k) = \left|S_w^a(\mathbf{x}, n, k)\right| Y(n, k) \tag{2}$$

where $\hat{Y}_w(\mathbf{x}, n, k)$ is the time-frequency signal at an arbitrary location, $\mathbf{x}$, in the reproduction region, $\mathbb{D}$, $Y(n, k)$ is obtained from the short-time Fourier transform of the windowed frame of input $y(n)$ and $|\cdot|$ denotes the absolute value. Using overlap-add reconstruction we can obtain the time-domain signal at any point in $\mathbb{D}$ where a different weighting function can be used for each space-time-frequency. The weighting function can now be used to control the leaked content into the quiet zone in the space-time-frequency domain.

III. Psychoacoustic Weighting Models

The capability of controlling the energy leakage between zones then allows the weighting function to become dependent on the signal being reproduced. For instance, the leaked audio spectrum may be controlled, altered, suppressed or designed to be masked by another spectrum. From this, psychoacoustic modelling can be applied to the weighting function in order to reduce the perceptual affect of the leakage in the quiet zone.

A. The Hearing Threshold

The benefit of using zone weighting is that the hard constraint of zero energy is alleviated and sound energy may be allowed to leak into the quiet zone. However, this then means the quiet zone is no longer completely quiet.

Due to the human threshold of hearing in quiet, a quiet zone could be redefined so that the sound pressure level is imperceptible. This would then make a weighted multizone system practical (from a relieved constraint) and remain quiet (perceptually). The threshold in quiet has been well established with frequency dependent functions that provide a good approximation [11], [13].

Using the new space-time-frequency domain weighting it is possible to apply the threshold in quiet approximation to (2) where $w(\mathbf{x}, n, k)$ is chosen so that the output in the quiet zone, $\hat{Y}_w(\mathbf{x}_q, n, k)$, is as close to the threshold in quiet as possible. Then, using the codebook method [10], $w(\mathbf{x}, n, k)$ can be chosen to minimise the difference,

$$\min_{w_q} \left(\hat{Y}_w(\mathbf{x}_q, n, k) - A(\mathbf{x}_q, n, k)\right) \tag{3}$$

where $A(\mathbf{x}_q, n, k)$ is a space-time-frequency dependent function describing the perceptual criteria. In this work Sound Pressure Level (SPL) in dB is relative to the threshold of hearing $p_r = 20\,\mu\mathrm{Pa}$.

B. Spreading Functions to Reduce Multizone Error

Analysing the weighted multizone reproductions in [7] reveals that larger weighting increases the error in the bright zone whilst suppressing the quiet zone. The quality of the spatial reproduction in the bright zone is less erroneous when the weighting is eased for the quiet zone allowing more energy to leak. If the quiet zone is controlled to have minimal energy leaked into it there becomes an erroneous bright zone.

The spatial errors shown in Fig. 2 are calculated from [7]:

$$\epsilon_b(n, k) = \frac{\int_{\mathbb{D}_b} \left|S^d(\mathbf{x}, n, k) - S_w^a(\mathbf{x}, n, k)\right|^2 d\mathbf{x}}{\int_{\mathbb{D}_b} \left|S^d(\mathbf{x}, n, k)\right|^2 d\mathbf{x}} \tag{4}$$

where $\epsilon_b(n, k)$ is the spatial error in the bright zone.

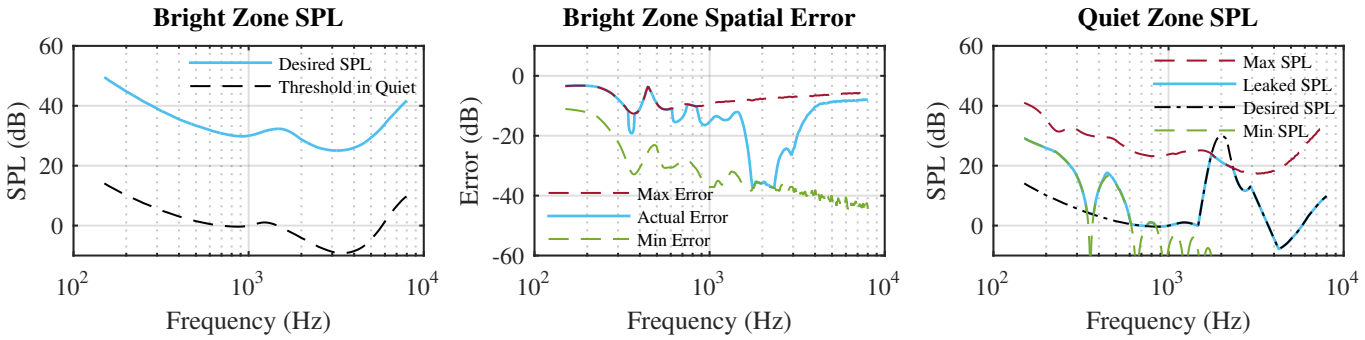

Fig. 2: Multizone soundfield reproduction with perceptual weighting in the quiet zone. The desired bright zone signal is an equal loudness curve at 30 phon [13] and a 2 kHz masker signal at 30 dB SPL is present in the quiet zone. The red and green dashed lines show the worst and best case scenarios, respectively. The bright zone error is calculated using (4). The "Leaked SPL" shows the result after controlling the interzone interference with w_q.

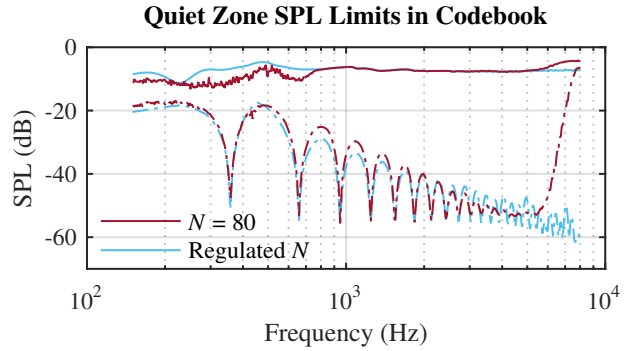

Fig. 3: Shows the maximum (solid line) and minimum (dash-dot line) levels in a codebook that the weighting function provides for $w_q = 10^{-2} \rightarrow 10^4$. The number of basis planewaves, N, used to generate the codebooks are a constant number, 80, (red lines) and a regulated number (blue lines).

The work in [7] shows that for $k = 2\,\text{kHz}$ the spatial error is greater than $-5\,\text{dB}$ when the quiet zone is occluded by the bright zone and has a large weight (equivalent to $w_q = 10$), however, the spatial error is less than $-20\,\text{dB}$ when the weight is alleviated (equivalent to $w_q = 0.1$).

In Fig. 2 it is shown that using a spreading function to mask apparent sounds, in the "target" quiet zone, can reduce the error of the reproduction in the bright zone. This is because we can safely allow the sound energy at particular frequencies to leak into the quiet zone with no perceptual affect. If the "target" quiet zone contained many different frequency components then it is possible that the bright zone energy could be completely leaked into the quiet zone unperceivably and thus reduce the error in the bright zone to a minimum.

C. Maintaining Broad Control for Wideband Speech

Inaccurate reproductions can be caused by spatial aliasing and poorly conditioned matrices which cause the control range of the weighting function per frequency to become reduced and less accurate as can be seen in Fig. 3. In order to maintain accurate estimation of the affect of different zone weights the number of basis planewaves needed to reproduce the soundfields can be regulated [10].

Fig. 3 shows that with a regulated N the difference in level that a given w_q can provide in the quiet zone is improved for low and high frequencies and the response is smoother for low frequencies. $N = 80$ is well balanced between spatial aliasing and ill-conditioning for 2 kHz [7], [10].

IV. Results

A. Multizone Reproduction Evaluation Setup

The multizone soundfield layout of Fig. 1 is evaluated, where $r = 0.3\,\text{m}$, $r_z = 0.6\,\text{m}$, $R = 1\,\text{m}$, $R_l = 1.5\,\text{m}$, $\theta = \sin^{-1}(r/2r_z) \approx 14.5°$ and $\pi \approx 3.141\,59\,\text{rad}$. The value of θ is chosen such that an evanescent planewave with instant decay would interfere with half the quiet zone [7], [10]. This choice results in a slight occlusion problem where the range of weighting control is larger than for no occlusion and full occlusion. Signals sampled at 16 kHz are converted to the time-frequency domain using a Hamming window (50% overlap) and Fast Fourier transform (FFT) of length 1024. For the evaluation we use the efficient method of codebooks described in [10] to store pre-determined weighted soundfield values to be used for a given setup or wideband reproduction. The codebooks are constructed for a reproduction where $L = 65$ and $\phi_L = 2\pi$ which for this particular setup is free of significant aliasing problems in the quiet zone below approximately 8 kHz.

The codebooks are built with spatial pressure samples for all $\mathbf{x} \in \mathbb{D}_b \cap \mathbb{D}_q$ with each soundfield zone approximated from 2724 samples. The zone weights are chosen as $w_b = 1$ and $w_u = 0.05$ following [7], [10] and the variable weight is w_q.

Then, using (3), w_q is chosen to match the quiet zone to a given level, $A(\mathbf{x}_q, n, k)$. In this work we choose $A(\mathbf{x}_q, n, k)$ to be the threshold in quiet using the ISO226 standard [13] with additional masking curves using the ISO/IEC MPEG Psychoacoustic Model 2 spreading function [11].

Speech files for the evaluation are taken from the TIMIT corpus [14] where 20 files are chosen randomly. The male to female speaker ratio of these files is 50 : 50. Evaluations using these speech files are shown with 95% confidence intervals.

B. Reduced Bright Zone Error from Psychoacoustic Masking

The error induced from the multizone reproduction of the speech soundfields is evaluated using the Mean Squared Error (MSE) of the reproduced speech where the reference signal for the MSE is the original speech signal. To obtain an approximation of the reproduced speech the mean of the

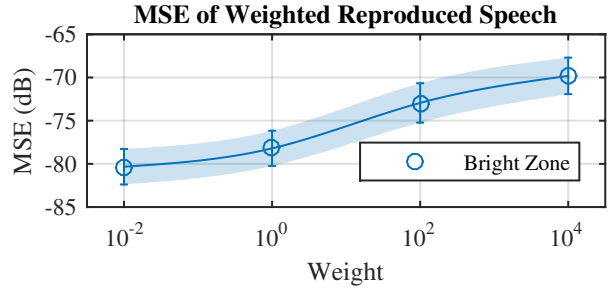

Fig. 4: Shows the MSE of reproduced speech files in the bright zone for different uniform weighting functions (w_q).

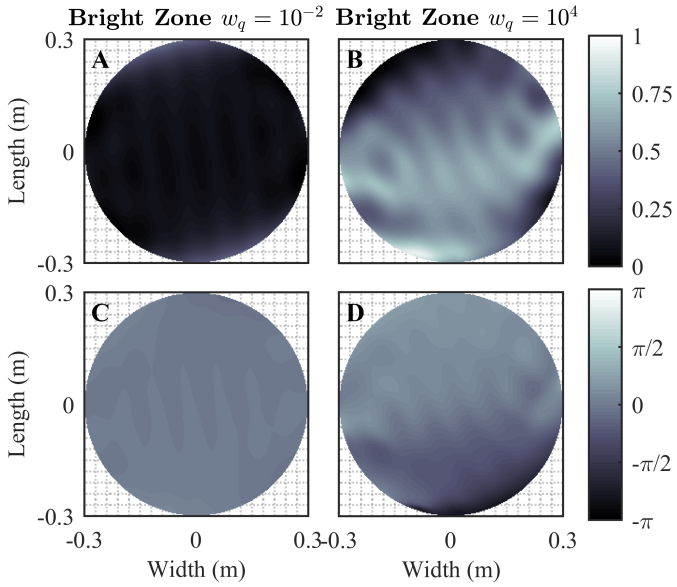

Fig. 5: Difference between $S^d(\mathbf{x}_b, n, k)$ and $S^a_w(\mathbf{x}_b, n, k)$ for $f = 2\,\text{kHz}$. **A** and **B** show the magnitude difference and **C** and **D** show the phase difference. **A** and **C** are for $w_q = 10^{-2}$ and **B** and **D** are for $w_q = 10^4$.

simulated spatial pressure samples obtained with the approach of section III are used across $\mathbb{D}_b$ and $\mathbb{D}_q$.

Upon analysing the MSE of different reproduced speech files it becomes apparent that the majority of error measured in the bright zone from the reproduction is in the spatial domain. The sampling theory used to obtain the reproduced speech means that spatial information is neglected, however, (4) evaluates the spatial error and is similar to the measure of planarity [6]. This then means that the application of perceptual criteria primarily reduces the spatial error of the multizone reproduction.

The maximum improvement in MSE of the bright zone reproduced speech is $-10.5\,\text{dB}$, from $-69.8\,\text{dB}$ for $w_q = 10^4$ to $-80.3\,\text{dB}$ for $w_q = 10^{-2}$, and can be seen in Fig. 4. Even though there is a difference of $-10.5\,\text{dB}$, the MSE in the reproduced speech is minimal. However, the maximum improvement in spatial error for the bright zone, ϵ_b, averaged for all frequencies is $-24.0\,\text{dB}$, from $-7.4\,\text{dB}$ for $w_q = 10^4$ to $-31.5\,\text{dB}$ for $w_q = 10^{-2}$, and can be seen in Fig. 2.

Also shown in Fig. 2 is that a $2\,\text{kHz}$ masker signal in the quite zone can allow the spatial error in the bright zone to be reduced. This reduction in spatial error is depicted in Fig. 5 where the perceptual weighting uses $w_q = 10^{-2}$ instead of $w_q = 10^4$ which gives a smaller difference between the desired soundfield and reproduced soundfield. In Fig. 5 the magnitude difference is calculated from $|S^d| - |S^a_w|$ and the phase difference from $\arg(S^d/S^a_w)$. The equivalent improvement in ϵ_b and required loudspeaker power due to the perceptual weighting is $-28\,\text{dB}$ and 65% less, respectively.

V. Conclusions

In this paper we have proposed a method for perceptually weighting multizone speech soundfields which can improve error in bright zones, especially when the occlusion problem is apparent. We have shown the need for regulating the number of basis planewaves used for the reproduction. Perceptual weighting is shown to improve the MSE for reproduced speech in the bright zone from $-69.8\,\text{dB}$ to $-80.3\,\text{dB}$ and significantly reduce the spatial error on average from $-7.4\,\text{dB}$ to $-31.5\,\text{dB}$ whilst requiring less power. Future work includes testing methods for maximising the speech intelligibility difference and privacy between zones in multizone speech soundfields.

References

[1] M. Poletti, "An Investigation of 2-D Multizone Surround Sound Systems," in *Audio Engineering Society Convention 125*. Audio Engineering Society, Oct. 2008.

[2] W. Jin and W. Kleijn, "Multizone soundfield reproduction in reverberant rooms using compressed sensing techniques," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4728–4732.

[3] T. Betlehem, W. Zhang, M. Poletti, and T. D. Abhayapala, "Personal Sound Zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 81–91, 2015.

[4] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1711–1720, 2011.

[5] T. Betlehem and P. D. Teal, "A constrained optimization approach for multi-zone surround sound," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 437–440.

[6] P. Coleman, P. J. Jackson, M. Olik, and J. A. Pedersen, "Personal audio with a planar bright zone," vol. 136, no. 4, pp. 1725–1735.

[7] W. Jin, W. B. Kleijn, and D. Virette, "Multizone soundfield reproduction using orthogonal basis expansion," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 311–315.

[8] H. Chen, T. D. Abhayapala, and W. Zhang, "Enhanced sound field reproduction within prioritized control region," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 249. Institute of Noise Control Engineering, 2014, pp. 4055–4064.

[9] N. Radmanesh and I. S. Burnett, "Generation of isolated wideband sound fields using a combined two-stage lasso-ls algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 378–387, 2013.

[10] J. Donley and C. Ritz, "An efficient approach to dynamically weighted multizone wideband reproduction of speech soundfields," in *China Summit & International Conference on Signal and Information Processing (ChinaSIP)*. IEEE, 2015, pp. 60–64.

[11] M. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards*. Springer, 2003.

[12] K. Baykaner, P. Coleman, R. Mason, P. J. Jackson, J. Francombe, M. Olik, and S. Bech, "The relationship between target quality and interference in sound zone," vol. 63, no. 1, pp. 78–89.

[13] *Acoustics–Normal Equal-Loudness-Level Contours*. International Standard ISO 226, 2003.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.