# Vowel Identification in Stationary Broadband Sources Under Variable Loading

Vera M. Solenti
*Speech Measurement Section*
*Altirena Institute of Acoustics*
Ferrara
Italy
v.solenti@altirena.example

Esben K. Haldvik
*Department of Electrical Measurement*
*Østerdal College of Engineering*
Aalborg
Denmark

Maëlle A. Dessery
*Speech Measurement Section*
*Altirena Institute of Acoustics*
Ferrara
Italy

capstract:2611.00024v1 [eess.AS] 12 Nov 2026

***Abstract*—We tested whether spectral templates from sustained human phonation predict the vowels listeners assign to stationary mechanical noise. The signal corpus comprised 18 domestic cylinder vacuum cleaners recorded in two rooms at four bag loads. Six machines supplied development material. Of 192 tokens from the remaining 12, 144 received a majority transcription from a panel choosing among three reference vowels and an outside category. A spectral-envelope classifier agreed with the majority on 113 of these tokens (78.5%). A linear-prediction baseline agreed on 87. The number transcribed as /u/ increased from 3 of 48 at empty bag to 19 of 48 at three-quarter load. In eight matched sweeps both endpoint transcriptions had majorities and changed from /a/ to /u/. Suppressing the narrow motor lines retained most of the agreement. Replacing the broadband envelope with its corpus mean largely removed it. The bag load affects the position of an appliance in a closed vowel inventory even during nominally constant operation.**

***Index Terms*—vowel identification, spectral envelope estimation, machine acoustics, listening tests**

## I. Introduction

The two lowest concentrations of energy in a sustained vowel account for much of its identity in isolation. A listener can often name the vowel when the excitation has little periodic structure. A difficulty arises when a recorder catches another source in the same frequency region. A recognizer can return a vowel for several seconds, with little frame-to-frame variation and no neighboring consonants. A duration test then gives considerable weight to a sound for which there was no speaker.

We encountered this case while scoring isolated Italian vowels in rooms occupied by cleaning staff. The disputed intervals lasted too long to belong to any of the elicited responses. Two transcribers separately entered /u/ during a passage in which the speaker’s lips were closed. The microphone had picked up a cylinder vacuum cleaner from the corridor. During a return visit, with a fresh bag in the same machine, one transcriber entered /a/. This observation left open whether the listener had supplied a vowel in response to the instructions or had identified a repeatable spectral feature. The original recordings have no bag record and contribute no observations to the study.

We treated the transcription as the outcome. The model sees a single level-matched excerpt of noise and selects one of four labels. Its vowel references all come from people. The evaluation asks whether the selection agrees with listeners, including those who assign an excerpt to the outside category. A separate analysis follows the same appliance through increasing loads. A vacuum cleaner whose transcriptions change between two bags still belongs to one test unit. All intervals for prediction error preserve that grouping.

## II. Acoustic basis

Peterson and Barney’s vowel measurements [1] give quite different distributions for neighboring vowels, with considerable variation among speakers within a vowel. The larger data set of Hillenbrand and colleagues [2] also contains temporal information which a sustained-state measurement discards. Italian listeners in our preliminary sessions could use the stationary portion of a reference token and repeat the label after a delay. We restricted the reference set to /a/, /i/ and /u/. Close and open mid vowels would require additional reference tokens for the regional varieties represented in the panel [3]. We use the symbols as names for response keys throughout.

Source-filter accounts of speech distinguish the excitation from the response of the tract [4], [5]. A fan has a broadband component with a superposed set of lines. The spacing of the lines depends on the rotating assembly. Air drawn through the head, hose and dust bag follows a different path from air radiated through a panel of the housing. A microphone receives a sum of these contributions [6], [7]. Hence there need be no pair of physical cavities corresponding to two peaks in the recorded spectrum. A measurement at one position can support a statement about vowel identity even when it supplies little information about the location of the resonances.

An all-pole fit gives a compact description of speech spectra [8], [9]. With a sufficiently strong narrow line it can spend a conjugate pole pair on the drive and shift one of the broader maxima across a vowel boundary. We also considered smooth spectra of the sort used for noise-excited phonetic stimuli [10]. In an auditory filter bank, neighboring high-frequency bins describe wider frequency intervals than neighboring low-frequency bins [11]. This limits the influence of individual fine spectral peaks. Our two classifiers share the vowel recordings, the portion of the signal admitted for analysis and the decision

rule. They differ in the construction of the spectrum from which the distance to a reference is measured.

## III. Material

### A. Units and load

We borrowed one specimen each of 18 bagged cylinder designs from two appliance collections. All ran at a fixed operator setting and allowed the same head to remain fitted throughout a sweep. We retained the stock hose and floor head belonging to each specimen. A design with a continuously varying automatic setting did not enter the inventory. The allocation to development or test took place before recording. We sorted by nominal electrical power, formed six consecutive groups of three and drew one machine from each group for development. The other two entered the test set. Nominal power served only to distribute the larger motors between sets.

The laboratory made four replacement bags for each specimen from one lot of its specified bag stock. A rigid gauge gave the expanded volume of a bag without tension on the seams. We divided loose dry cellulose by volume in that gauge. At load fractions 0, 0.25, 0.50 and 0.75, a retaining mesh held the material against the same rear surface. The mesh was present in the empty bag as well. These fractions describe the occupied volume under the preparation procedure. Dust retained in service has a different distribution of particle sizes and a bag bearing the same nominal fraction can have a different pressure drop [12]. At three-quarter volume every machine still ran at the fixed setting for a complete take.

Each machine occupied a marked position in a furnished listening room and in a smaller absorption-lined room. We used two full installations in each room. Between installations we disconnected and reseated the hose, removed the bag and reset the floor head against its stops. Each installation included the four load fractions in a separate randomized order. Thus a sweep has four tokens associated by appliance, room and seating, although the empty-bag token need not precede the loaded token on the recording. All load comparisons in Section VI respect this pairing.

### B. Acquisition

A free-field microphone faced the midpoint of the hose connection at one meter. The receiver position was specified relative to the cabinet after the handle had been folded down. Outlet orientation consequently varied among specimens. We did not rotate a machine to maximize a vowel response. Twenty seconds of operation preceded a six-second recording at 48 kHz. A script retained the middle second, applied a 20 ms cosine taper at each end and matched its mean-square level to that of the reference set. A second microphone at the exhaust served only for checking that the recorded interval contained steady operation. Its signal was absent from both the listening material and the recognizers.

The intake head rested on a frame with a perforated plate in place of a floor covering. We left the plate in place for all bags at both sites. A change of carpet pile can alter the spectrum at the ear far more than moving a microphone a few centimeters [13]. We recorded the frame position and rotated the order of rooms by specimen. Both installations of a given machine in a room were completed during the same visit. Day and room are consequently confounded for an individual specimen. Differences between rooms include setup on a different day.

Table I gives recording counts before listening. There were no automatic replacements of tokens on the basis of spectral shape. A failed recording could be repeated only when a missing file or digital overload was evident before the panel allocation. Two overloaded development takes were repeated at lower preamplifier gain. Every retained take passed the level check. We kept the original identifiers for the repeats and counted each retained take once.

TABLE I
Number of distinct units and retained one-second tokens. Each machine supplied four loads in each of four installations.

| Set | Machines | Installations | Tokens |
|---|---|---|---|
| Development | 6 | 24 | 96 |
| Test | 12 | 48 | 192 |
| Total | 18 | 72 | 288 |

### C. Human reference material

Nine Italian-speaking adults recorded six sustained tokens per vowel in the lined room. We removed the onset and release, leaving a central second. Each speaker contributed the same number of tokens even when several takes had nearly identical envelopes. First averaging within speaker prevented a speaker with exceptionally steady phonation from acquiring extra weight through the number of accepted frames. The three reference envelopes used all 162 tokens. None of these speakers served on the test panel. A separate group of 12 listeners transcribed all 96 development tokens using the same anchors and four response keys as the test listeners. Of these tokens, 70 had a majority transcription and entered cutoff selection.

## IV. Signal analysis

Spectra used 40 ms Hann windows at 10 ms spacing. We averaged the squared transform magnitudes across the central 800 ms [14]. At 48 kHz a frame contains 1,920 samples. Padding the transform to 2,048 points allowed a common frequency grid for every recording. We discarded frequencies outside 180 to 3,500 Hz after estimating the spectrum. The low cutoff avoids much of the motor fundamental and the upper cutoff admits the second concentration of /i/ in every reference speaker. A check with an upper limit of 2,500 Hz appears in Table II.

A running median in a 150 Hz window provided the local background at each frequency. For appliance tokens only, a value more than 10 dB above this background was replaced by the background estimate. The median and the decision both used the original spectrum, so a replaced bin could not change the status of its neighbors. This operation removes isolated lines and leaves a broad maximum wider than the window

largely intact. Applying it to human speech would delete part of the reference harmonic structure, particularly at high fundamental frequency. We therefore integrated the original human spectra in the same filter bank without line replacement.

Let $b_k$ be the logarithm of power in the $k$th of 32 overlapping triangular bands equally spaced on the equivalent rectangular bandwidth rate scale. We remove the mean across bands to form $v_k$. For vowel $j$ let $r_{jk}$ be the mean of the corresponding speaker averages. We used the mean squared discrepancy

$$D_j(v) = \frac{1}{32}\sum_{k=1}^{32}(v_k - r_{jk})^2. \tag{1}$$

Here the band powers entering the logarithm have a common floor 60 dB below the largest band power in the token. Both $v$ and $r$ are in dB, so the cutoff on Eq. (1) has units of squared dB. The smallest distance supplies the vowel label unless it exceeds a cutoff, in which case the prediction is outside. The cutoff maximized agreement with majority responses on the subset of the 96 development tokens with a majority transcription. Ties in the optimization went to the smaller cutoff. The selected value was 31 square dB. A single threshold served all three vowels. We retained it for both test rooms.

The linear prediction baseline used an autocorrelation fit of order 12 to a signal resampled to 8 kHz. We evaluated its smooth log power spectrum on the filter bank, repeated the reference processing with the same estimator, and selected its own rejection cutoff on development. A classifier based on the strongest narrow line saw only the frequency of that line. It assigned the majority development response in one of eight fixed log-frequency bins, with the overall development mode in any unoccupied bin. Such a classifier can exploit covariance between motor size and housing without any estimate of the outer spectral envelope. A constant classifier returned /a/, the most frequent development majority.

## V. Listening and inference

Each test token received one response from each of 12 listeners. There were 24 listeners in total. A balanced allocation gave each person 96 different test tokens, with 24 at each load. No person heard a test token twice in the main block. Half the panel heard one allocation and half its complement. We varied presentation order within allocation and prevented immediately successive items from the same machine. Listeners worked in isolation with the same headphone model. The display carried the three vowel symbols and a fourth key labeled “none of these.” The anchor bank contained one held vowel per symbol from each of three reference speakers. Listeners could revisit the bank between trials.

The consent form described a task on speech-like qualities of recorded sounds. The panel received no information about bag loads, appliance models or experimenters’ predictions. Several listeners recognized the source family within the first block. Masking source identity throughout an experiment of this length would require a different stimulus. Listeners gave written consent under the auditory perception protocol of the Altirena institute. We asked about prior transcription experience after the last block. Sixteen of the 24 had used the international phonetic alphabet professionally. Eight had received at least a semester of instruction. We did not recruit listeners on the basis of their performance on the mechanical recordings.

A majority requires seven identical responses including an outside response. A token split six to six between a vowel and outside has no majority. Pluralities with five or six votes also fail the criterion [15]. We report both the number of qualifying tokens and the fraction correctly predicted among those tokens. In addition, the vote share of the predicted label can be evaluated on every token. For $N$ tokens with 12 votes each, agreement with an individual vote is

$$A_{\mathrm{vote}} = \frac{1}{12N}\sum_{n=1}^{N} c_{n,\hat{j}_n}, \tag{2}$$

where $c_{n,j}$ counts votes for response $j$. Eq. (2) gives less weight to narrow majorities than to unanimous transcriptions. It has a different ceiling from agreement with the panel mode. Bootstrapping resampled the 12 test machines with replacement, carrying along an entire machine’s recordings and vote counts [16], [17]. We took the 2.5th and 97.5th percentiles of 10,000 resamples. These intervals describe variation over machines for these listener allocations. They leave the population of listeners fixed.

## VI. Results

A majority transcribed 144 of 192 test tokens (75%). Of the 144, 115 had a vowel majority and 29 had an outside majority. Fig. 1 gives the underlying counts. The envelope classifier correctly returned a vowel for 93 of the 115 vowel majorities and outside for 20 of the 29 outside majorities. Across the four classes it agreed on 113 of 144, or 78.5% (machine bootstrap interval 71.2 to 84.8%). /u/ had the highest recall. Its principal confusion with /a/ occurred in the furnished room, where two cabinets emitted substantial low-frequency energy from the side opposite the microphone.

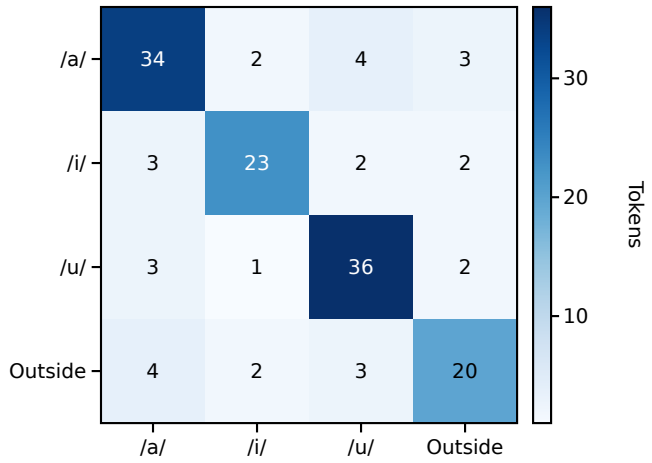



Fig. 1. Test tokens with a majority response. Rows give listener transcriptions and columns give the envelope prediction.

The gap from linear prediction was 26 tokens on the same denominator (Table II). In a paired resampling of whole machines the agreement difference was 18.1 percentage points (95% interval 9.3 to 26.9). The strongest-line classifier had little success outside the low-frequency bins. Ten of its correct decisions concerned three housings with a strong component just above the lower cutoff. In a transfer to a different collection the prevalence of those housings could readily change. Inspecting the test errors after the analysis suggested a second spectral maximum consistent with the final /u/ on two machines. That inspection played no part in selecting the cutoffs.

TABLE II
CORRECT CLASSIFICATIONS OF THE 144 MAJORITY TOKENS. CONTROLS USE THRESHOLDS SELECTED ON DEVELOPMENT FOR THEIR OWN PROCESSING.

| System | Correct | Agreement (%) |
|---|---|---|
| Development mode | 43 | 29.9 |
| Strongest line | 54 | 37.5 |
| Linear prediction | 87 | 60.4 |
| Smooth envelope | 113 | 78.5 |
| Smooth envelope, 2.5 kHz ceiling | 103 | 71.5 |
| Original lines retained | 106 | 73.6 |
| Common envelope, original lines | 57 | 39.6 |

On all 192 tokens the predicted labels received 1,402 of 2,304 votes (60.9%). Forty-eight tokens contribute to this number which cannot enter the usual confusion matrix. For an item with a four-way division of votes, a reasonable top label still gathers only about a quarter of the individual responses. The ratio in Eq. (2) thus answers the question relevant to a single transcriber asked to annotate a passage alone. The larger value in Table II applies when a convened panel can leave a passage unresolved.

### *A. Load changes within a machine*

At empty bag, 18 of 48 test installations yielded /a/ and three yielded /u/. At three-quarter load these counts were four and 19. The intermediate loads in Fig. 2 show a progressive shift in the margins. Each column represents the same installations. A column includes tokens without a majority, so differences cannot arise from a smaller denominator at high load. For /i/ the counts stayed between six and eight across the four settings. The outside response increased slightly at the highest load.

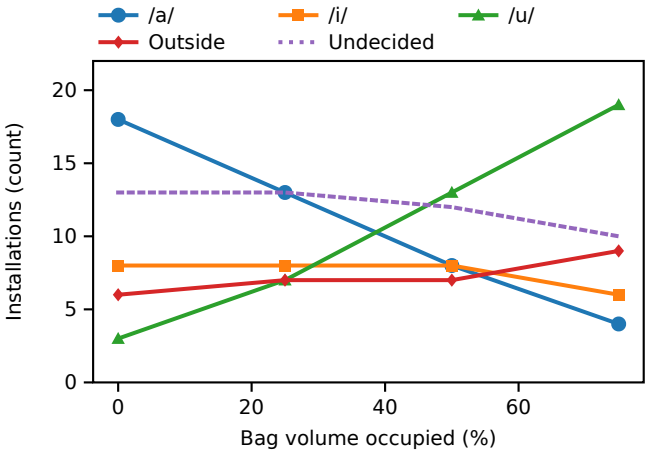



Fig. 2. Majority transcriptions by occupied bag volume. Each load includes 48 test installations.

The paired records put a limit on the linguistic interpretation of those margins. Twenty-nine installations had majorities at both endpoints. Eight of these went from /a/ to /u/, and none went from /u/ to /a/. Eleven retained the same endpoint transcription, including three with outside at both ends. The remaining ten changed between other ordered pairs of responses. The eight open-to-close changes came from five distinct machines. Treating them as eight independent changes would count several responses to the same cabinet twice. We did not use the intermediate states to assign a conventional diphthong. They came from separate takes and their physical order was randomized.

Fig. 3 shows the empty and three-quarter spectra for a machine with a majority switch in all four installations. The first broad maximum shifts down and the upper shoulder weakens. Narrow lines in its unprocessed signal move slightly upward under load. A bank of references indexed by motor frequency would put these observations in the opposite order from a classifier based on the location of the lower broad maximum. There is enough airflow noise between the lines to estimate that maximum without placing a harmonic there.

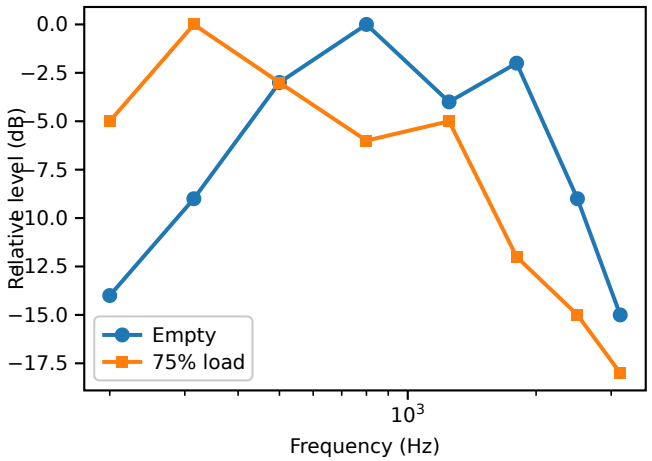



Fig. 3. Band levels relative to the strongest displayed band for specimen T07 in the lined room, first seating. Values are local averages after line suppression.

### B. What carries the vowel

For the common-envelope control we divided each power spectrum by its smoothed envelope and multiplied the residual by the geometric mean of the development appliance envelopes. A floor on the denominator limited correction in the deep spectral troughs. The test signal retained the frequencies and relative heights of its narrow lines above the common envelope. Distances and thresholds for this control were measured using the resulting spectra without a second line removal. The loading effect was much smaller than in the main analysis. On the 48 three-quarter tokens the control predicted /u/ on seven, compared with 21 for the ordinary envelope classifier. These are classifier outputs for all tokens and include ones with no majority.

We examined whether the recorded room could supply a vowel without the machine. The transfer characteristic measured from the cabinet mark to the receiving position can have several narrow dips at the listening-room microphone. We played a common, spectrally flat noise signal from a small loudspeaker at the cabinet mark, four times in each room. All eight recordings fell outside the rejection surface. Moving the loudspeaker to the far end of the floor head produced one /a/ in four furnished-room repeats. A distributed source in such a room may therefore acquire some of its apparent articulation through the measurement geometry. The single-speaker control cannot reproduce this distribution.

## VII. Interpretation

A listener transcribing a stationary vacuum cleaner can distinguish three vowel qualities under a fixed set of alternatives. Only a subset of the recordings supports a reproducible positive label. Bag loading moves the distribution toward /u/ in this collection, and individual machines make the same change after disassembly. A recognition system using only human sustained vowels as its positive templates would therefore benefit from a separate test for speech presence before treating a long /u/ as a held human response. Increasing the minimum duration alone would admit every central second in this corpus.

Mechanical explanations remain underdetermined. Packing the rear wall can alter acoustic loss below 1 kHz, the fan operating point and the fraction of microphone power that arrives from the outlet. The broad spectra reported here cannot distinguish those contributions [18]. A transfer measurement across the bag would require an acoustic port that the collections did not permit us to fit. We could vary the bag while leaving the hose geometry approximately fixed. This intervention establishes a load effect for the assembled machines under the plate and ram pressure used here. It does not identify an anatomical equivalent for the lower formant.

We deliberately did not ask listeners for words. A sequence that begins /a/ and ends /u/ in Fig. 2 took the operator four bag changes to produce and has no consonantal boundaries. It cannot be transcribed as connected speech on these data. Work on perceptual organization gives other examples in which sparse cues support a speech percept under instructions [19], [20]. Here the outside key and the explicit accounting of divided votes prove consequential. Forcing the closest vowel on every token would add 77 vowel transcriptions to a table in which only 115 have majority support.

The collectors supplied relatively serviceable machines with replaceable bags. Bagless designs admit a different acoustic path, and a large portion of the portable market has no hose. The present reference space also omits mid vowels. A motor that appeared inconsistent between /i/ and /u/ might have drawn a consistent fourth response on another inventory. Agreement with an instructed transcription panel can travel across rooms more easily than the underlying interpretation travels across languages. Reuse of these labels requires the anchors and response sheet as well as the audio.

## VIII. Conclusion

For archival annotation we retain an appliance identifier with each mechanical vowel. On machine T07 an annotator receiving only the last bag state would enter /u/ against all four installations. The same identifier is needed to recover the four empty-bag /a/ entries from the earlier sheet. Splitting the records by vowel before constructing a test set would place the same motor, fan and hose in both partitions.

## Appendix A. Token accounting

Table III gives counts behind Fig. 2, including unresolved trials. The apparent shift would be exaggerated by discarding undecided trials separately at each endpoint. For example, the high-load fraction of /u/ among majority trials is 19/38, while the corresponding fraction across all installations is 19/48. The latter fraction answers the deployment question about how many recordings would actually receive an agreed /u/. The paired calculation in Section VI additionally requires a resolved empty-bag record for the same seating.

TABLE III
TRANSCRIPTION COUNTS USED IN THE LOAD ANALYSIS. EACH ROW INCLUDES ALL 48 INSTALLATIONS.

| Load (%) | /a/ | /i/ | /u/ | Outside | Undecided |
|---|---|---|---|---|---|
| 0 | 18 | 8 | 3 | 6 | 13 |
| 25 | 13 | 8 | 7 | 7 | 13 |
| 50 | 8 | 8 | 13 | 7 | 12 |
| 75 | 4 | 6 | 19 | 9 | 10 |

## APPENDIX B. INTERPOLATION AND LEVEL CONVENTIONS

For an end band the first or last triangular half-window terminates at the analysis cutoff. We divided each weighted sum by the sum of its weights before taking decibels. This prevents the unequal widths of the bands from assigning an upward tilt to a flat power spectrum. Repeating the reference analysis with unnormalized triangular sums changes the shape of every displayed envelope. When the same term is added to both vectors in Eq. (1) it cancels after mean subtraction. A discrepancy can remain at the clipping floor, which depends on the token’s largest band.

The spectra in Fig. 3 are normalized to their own largest displayed band for legibility. Classification always uses all 32 bands with the mean subtracted. We stored those vectors before making the figure. Deriving a distance by reading back the eight plotted values would omit the high-frequency shoulder that separates several /i/ majorities from outside.

## ACKNOWLEDGMENTS

Sofia Velardi kept the preparation and listening sheets in separate custody until the transcriptions were complete. The two collections supplied bags from their reserve stock.

## REFERENCES

[1] G. E. Peterson and H. L. Barney, “Control Methods Used in a Study of the Vowels,” *Journal of the Acoustical Society of America*, vol. 24, pp. 175–184, 1952.

[2] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, “Acoustic characteristics of American English vowels,” *Journal of the Acoustical Society of America*, vol. 97, pp. 3099–3111, 1995.

[3] L. C. Serandi and A. Volnès, “Regional response boundaries in an isolated mid-vowel task,” *Quaderni di Fonologia Sperimentale*, vol. 11, pp. 43–61, 2018.

[4] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1960.

[5] K. N. Stevens, *Acoustic Phonetics*. MIT Press, 1998.

[6] T. E. Haldrin, “Coherent and diffuse contributions to the radiation of an enclosed blower,” *Transactions on Enclosure Acoustics*, vol. 16, pp. 79–96, 2019.

[7] D. F. Porelli and A. Marest, “Source substitution in domestic suction equipment,” *Annals of Applied Sound Measurement*, vol. 7, pp. 13–29, 2021.

[8] B. S. Atal and S. L. Hanauer, “Speech Analysis and Synthesis by Linear Prediction of the Speech Wave,” *Journal of the Acoustical Society of America*, vol. 50, pp. 637–655, 1971.

[9] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, pp. 561–580, 1975.

[10] O. M. Berris, “Noise excitation and the identification of incomplete vowel spectra,” *Studies in Auditory Transcription*, vol. 9, pp. 201–217, 2016.

[11] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hearing Research*, vol. 47, pp. 103–138, 1990.

[12] I. V. Mirel, “Volume and resistance as independent load measures for porous collectors,” *Bulletin of Flow Measurement Practice*, vol. 14, pp. 82–98, 2020.

[13] H. Orvaldi and L. P. Sénac, “Receiver repeatability for a movable intake above textile surfaces,” *Measurement Notes in Room Acoustics*, vol. 5, pp. 27–39, 2017.

[14] P. D. Welch, “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, 1967.

[15] F. Daurel and L. A. Svevik, “Majority and outside responses in mechanical sound sorting,” *Quarterly of Experimental Listening*, vol. 18, pp. 52–69, 2023.

[16] B. Efron, “Bootstrap Methods: Another Look at the Jackknife,” *Annals of Statistics*, vol. 7, pp. 1–26, 1979.

[17] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*. Cambridge University Press, 1997.

[18] G. Neraux and E. Korrali, “Inverse ambiguity in two-port measurements with a distributed radiating case,” *Cahiers de Métrologie Sonore*, vol. 20, pp. 107–126, 2022.

[19] R. E. Remez, P. E. Rubin, D. B. Pisoni, and T. D. Carrell, “Speech perception without traditional speech cues,” *Science*, vol. 212, pp. 947–949, 1981.

[20] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.