Estimation of multiple-branch vocal tract models: the influence of prior assumptions

Christian H. Kasess[1], *Wolfgang Kreuzer*[1]

[1]Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

christian.kasess@oeaw.ac.at, wolfgang.kreuzer@oeaw.ac.at

Abstract

Branched-tube models can be used for modeling nasal speech such as nasal stops and nasalized vowels. Previously, it has been shown that the use of probabilistic prior information such as smoothness priors can reduce the within-speaker variability of the vocal tract estimates. This model, however, lacked a representation of paranasal cavities and thus a model with a more complex branching structure is desirable. This raises the question of what prior information is necessary for physically plausible parameter estimates. Here, a model with one maxillary sinus is estimated. The sinus is parameterized in terms of its resonance using radius and angle in the z-plane, and the coupling area ratio. The probabilistic scheme mentioned above is used to estimate nasal stops /m/ and /n/ extracted from the TIMIT database. Different prior assumptions concerning resonance frequency, bandwidth, and coupling of the sinus to the nasal cavity are tested. Results show, on average, a better model fit for the model including the sinus. Further, prior assumptions are shown to have a large influence on the estimated resonance of the sinus. In particular, the lack of anatomically motivated assumptions about the bandwidth and/or the resonance frequency yields unrealistic estimates of these values.

Index Terms: nasal stops, paranasal cavities, parameter estimation, prior information

1. Introduction

The human vocal tract (VT) can be roughly understood as a nonuniform acoustic tube terminated by the glottis and the lips. At the velum the nasal tract is coupled to the vocal tract. Modulation of the cross section by velum, lips, jaw, and tongue produces different sounds. As a consequence, branched-tube speech production models are an appropriate representation for speech when the velum is lowered, e.g. nasal stops and nasalized vowels. These computational models have been investigated in some detail [1, 2], however the estimation of such models from speech signals still proves problematic due to the complex relation between the poles and the zeros in the transfer function of the model.

Recently, a probabilistic estimation method for a branched-tube model was introduced [3] that allows to include probabilistic prior information about the vocal tract in order to regularize the solution. Further, contrary to previous methods [2, 4], this new method estimates the VT parameters based on the spectral envelope without requiring an explicit estimation of a pole-zero model. The estimation method was used to estimate a vocal tract for nasal stops as introduced in [4] using a branched-tube model consisting of three segmented tubes representing the pharyngeal, oral, and nasal cavities. It was shown that the use of priors that enforce smoothness of the vocal tract can reduce the within-speaker variability across VT area function estimates without strongly affecting the estimation error. The three-tube model used, however, lacks a representation of the paranasal cavities, i.e. the maxillary, frontal, and sphenoidal sinuses which are thought to be important structures for the transmission characteristics of the nasal tract (see e.g. [5]). Hence, more elaborate models are required in order to get a better representation of the spectral envelope.

One way is to use models with multiple branching points allowing for a much more realistic model of the nasal cavity that includes paranasal cavities as well as the left and right nostrils. Such multi-branch models can be readily derived from basic three-tube units (see e.g. [2]) as long as branching points are limited to three joining tubes which is sufficient in most cases. Applying these models, however, significantly increases the complexity of the estimation. In particular, the question about which prior information to include in order to yield meaningful estimates of the vocal tract parameters seems to be even more pressing as there are many potential physical limitations, e.g. the small coupling area of the paranasal cavities compared to the nasal area [6, 7] or restrictions on the resonance frequencies [5, 7], that may need to be considered.

The aim of this study is to investigate the effect of prior assumptions about the paranasal cavities on the estimation of the acoustic parameters of these cavities and on the overall modeling error. To this end, different assumptions about resonance frequency and bandwidth of a single paranasal cavity and the coupling to the nasal cavity were adopted and these models were estimated using the model inversion scheme introduced in [3].

2. Methods

2.1. Three tubes

The most basic branched-tube model for nasal speech consists of three tubes having L, M, and N segments representing the pharyngeal, nasal, and oral cavity, respectively. The three tubes are coupled at the velar junction. Using continuity conditions between the segments and at the coupling of the three branches, the forward volume velocity at the glottis u_g^+ is given in terms of the constant flow at the nostrils u_n as

$$\begin{aligned} u_g^+ &= \frac{\begin{pmatrix} 1 & \mu_{M+L} \end{pmatrix} \begin{pmatrix} D^+_{M+L-1}(z) \\ D^-_{M+L-1}(z) \end{pmatrix}}{C^+_{N-1}(z) + C^-_{N-1}(z)} u_n \\ &= E^+_{M+L-1}(z) + \mu_{M+L} E^-_{M+L-1}(z), \end{aligned} \quad (1)$$

with $E^\pm_{M+L-1}(z)$ representing the forward and backward travelling wave at the glottal end of the pharyngeal tube. The trans-

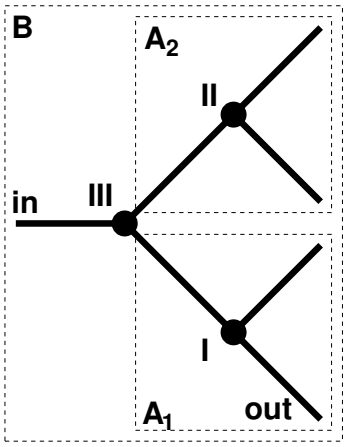

Figure 1: *Schematic representation of a multi-branch model consisting of three three-tube models A_1, A_2, and B. The roman numerals denote the different branching points.*

fer function is simply defined as $H(z) = u_n/u_g^+$. The polynomials $D^+_{M+L-1}(z)$ and $D^-_{M+L-1}(z)$ are given as

$$\begin{pmatrix} D^+_{M+L-1} \\ D^-_{M+L-1} \end{pmatrix} = \prod_{k=M+L-1}^{M} \begin{pmatrix} 1 & \mu_k \\ \mu_k z^{-1} & z^{-1} \end{pmatrix} \cdot \begin{pmatrix} P(z) & Q(z) \\ R(z) & S(z) \end{pmatrix} \cdot \prod_{l=M-1}^{1} \begin{pmatrix} 1 & \mu_l \\ \mu_l z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ z^{-1} \end{pmatrix} \quad (2)$$

and are of degree $M+L+N-1$ and $M+L+N$, respectively. The forward and backward components $C^{\pm}_{N-1}(z)$ in the oral cavity at the velum are defined as

$$\begin{pmatrix} C^+_{N-1} \\ C^-_{N-1} \end{pmatrix} = \prod_{k=N-1}^{1} \begin{pmatrix} 1 & \tilde{\mu}_k \\ \tilde{\mu}_k z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\mu}_0 z^{-1} \end{pmatrix}. \quad (3)$$

The $\tilde{\mu}_1, \ldots, \tilde{\mu}_{N-1}$ are the reflection coefficients for the oral part, $\mu_1 \ldots, \mu_{M-1}$ are the reflection coefficients for the nasal part and $\mu_M \ldots, \mu_{M+L}$ the reflection coefficients for the pharyngeal cavity. The reflection coefficient $\tilde{\mu}_0$ acts as an additional damping term at the terminal end of the oral tract that allows to model losses, i.e. damping effects at the lips [4]. For easier notation, scaling factors and the delay factor (see e.g. [8]) have been dropped in Eq. 2 as they do not influence the position of zeros and poles as long as there is only one point of emission. Models with multiple points of emission are not considered here. The coupling matrix between oral and nasopharyngeal tract is given as

$$\begin{aligned} P(z) &= C^+_{N-1} + (1-\nu)C^-_{N-1}, \\ Q(z) &= +\nu C^+_{N-1}, \\ R(z) &= -\nu C^-_{N-1}, \\ S(z) &= (1-\nu)C^+_{N-1} + C^-_{N-1}. \end{aligned} \quad (4)$$

Here, $\nu = \frac{\tilde{A}_{N-1}}{A_{M-1}+\tilde{A}_{N-1}}$ defines the relation of cross section areas of oral and nasal tract areas at the velum.

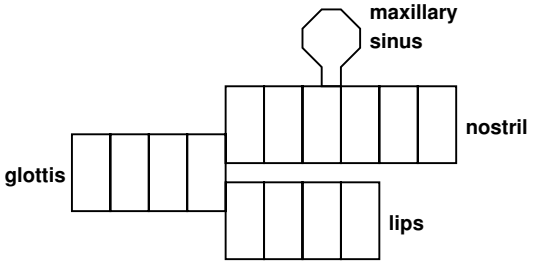

Figure 2: *Model structure of the multi-branch model used in this study. The model uses a maxillary sinus coupled to the main nasal cavity.*

2.2. General model

The model described in section 2.1 can be seen as a basic building block for a more general model including multiple three-tube junctions. The process of building this general model is relatively straightforward and here only a conceptual overview will be given using an example shown in Fig. 1 that could, for instance, describe the nasal cavity having left and right nostril with maxillary sinuses (branches A_1 and A_2) with the input at the velum. For a detailed derivation see e.g. [2].

First the branches A_1 and A_2 are calculated independently using the method described above yielding the flow at the left side at both branches in both directions $E^{\pm}_{A_1}(z)$ and $E^{\pm}_{A_2}(z)$, which are now rational functions. Afterwards, we use the three branch coupling procedure to combine the branches A_1, A_2, and B using Eqs. 1 to 4, where the right most product in Eq. 2 is replaced by $(E^+_{A_1}(z), E^-_{A_1}(z))^T$ and the denominator of the global transfer function (the analog to $C^+_{N-1} + C^-_{N-1}$ in Eq. 1) is replaced by the flow $E^+_{A_2}(z) + E^-_{A_2}(z)$ at the left end of branch A_2. So instead of simple polynomials the coupling matrix now combines rational polynomials. In the above example we define the transfer function as the ratio between the flow at the right lower end of branch A_1 and the flow at the left end of branch B, i.e. branch A_1 belongs to the main tract. As the rational polynomials $E^{\pm}_{A_1}(z)$ are the input on the right side of the "global" coupling matrix the roots of the denominator of $E^{\pm}_{A_1}(z)$ (i.e. the roots of $C^+_{A_1}(z) + C^-_{A_1}(z)$) are preserved (Eqs. 1 and 2). However, in the case of branch A_2 which acts as a side tract (see Fig. 1) the zeros of the rational polynomial $E^+_{A_2}(z) + E^-_{A_2}(z)$, i.e. the resonances (=poles) of branch A_2, become the zeros of the global tract. The denominator polynomial of $E^{\pm}_{A_2}(z)$ that represents the anti-resonances of branch A_2 simply vanishes as it factors out in both, the numerator and the denominator of the global transfer function (see Eqs. 1 to 4). As a consequence, the anti-resonances (i.e. zeros) of the main tract are preserved whereas the anti-resonances of the side tract can no longer be observed directly. They are, however, encoded in the resonances of the side tract. Of course, if e.g. the side tract is a simple tube, as will be the case here (Fig. 2), the side tract does not have any anti-resonances. Here, we only consider the case where tract A_2 is closed and hence does not emit sound. If A_2 is also open, the transfer function between the input and the emitting end of A_2 also needs to be calculated with the roles of A_1 and A_2 reversed.

2.3. Paranasal cavities

The paranasal cavities usually have a small coupling area and a comparatively large volume. Thus, a Helmholtz resonator

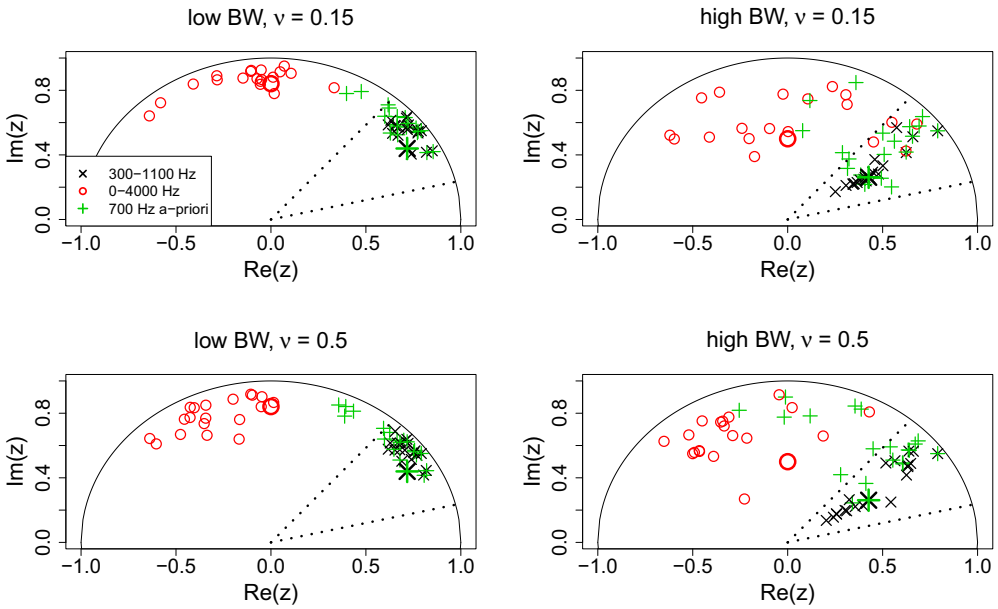

Figure 3: *Location of the estimated poles of the sinus for a single speaker. The colors code the assumptions about the resonance frequency whereas the different panels show different assumptions about coupling area ratio ν and bandwidth BW (left panels show low prior bandwidths, top panel shows small coupling areas of the sinus). The dotted lines mark 300 and 1100 Hz and the large symbols show the prior mean of frequency and bandwidth.*

(a structure with a single resonance frequency) is a commonly used model [5]. For the maxillary sinus for instance, this frequency has been estimated to lie in the range of approximately 300 to 1000 Hz [5] even though estimates as low as 100 Hz and as high as 2000 Hz exist [5, 7]. Estimates of the bandwidth of these resonators are reported to lie in a range of 100 to 200 Hz [5].

Clearly, such a structure with one resonance frequency of a certain bandwidth can most easily be incorporated into the multi-branch model described above as a segmented tube consisting of three segments. The parameterization in terms of reflection coefficients or areas, however, is problematic as it does not allow a simple control over the fundamental parameters of the paranasal cavity, i.e. resonance frequency and bandwidth of the resonator. To allow restrictions in the resonance frequencies of the cavities it is thus more useful to parameterize the cavity in terms of its radius and angle of the complex conjugate pole pair in the pole-zero plane. This allows a simple restriction to a certain frequency range and bandwidth using e.g. values from the literature. The polynomial generated by the paranasal pole pair is transformed back into a three-segment tube [9] and the coupling area that influences the effect of the resonator on the poles of the vocal tract is estimated in terms of ν i.e. the ratio of the paranasal area to the total area at the branching point.

2.4. Estimation scheme

The details of the estimation scheme have been described previously in [3]. Briefly, the estimation scheme fits the logarithm of the transfer function to the log of an estimate of the spectral envelope [10] based on the estimated pitch [11]. The estimation is based on a variational Bayesian scheme that allows to include a-priori assumptions about the vocal tract parameters. Prior information is included as Gaussian distributions, e.g. a smoothness prior for a reflection coefficient is centered around zero with a certain variance defining the strength of the assumption. Gaussian priors imply that the model parameters are defined on the whole real line. As the reflection coefficients are limited to the interval $(-1, 1)$, the estimated parameters are a transformed version of the true reflection coefficient. The transformation is a simple sigmoidal mapping that maps the model parameters that are defined on $(-\infty, \infty)$ onto $(-1, 1)$. Integrals in the estimation scheme are computed using the unscented transform [12] with a higher number of points to achieve a better integration.

2.5. A-priori assumptions

In [3] it was shown that mild smoothness assumptions about the vocal tract can decrease the within-speaker variance of the parameter estimates without strongly affecting the model error. It was found that a prior variance of $\sigma^2 = 0.1$ was the best setting and thus we will use this value for all experiments in this study The objective of this study was to test the effect of different assumptions on the estimation of the resonances of the paranasal cavities. Thus, combinations of different restrictions are tested for a model containing a model of the maxillary sinus as a side branch to the nasal tract (Fig. 2) located at an anatomically motivated postition [5, 6, 7] although it is important to note that the segment length of about 2 cm at 8000 Hz does not allow a very precise placement. First, the resonance frequency of the sinus is either restricted to lie between 300 and 1100 Hz [5] or to vary between 0 and 4 kHz. For the latter case, the prior expectation is either centered on the range (2 kHz) or set to the mean of the restricted estimation (700 Hz). Second, the prior expectation of the bandwidth is either set such that the radius of the resonance of the sinus in the pole-zero plane is 0.5 (bandwidth of roughly 1760 Hz) or such that the a-priori bandwidth amounts to a fourth of that value (440 Hz) which is a cautious assumption

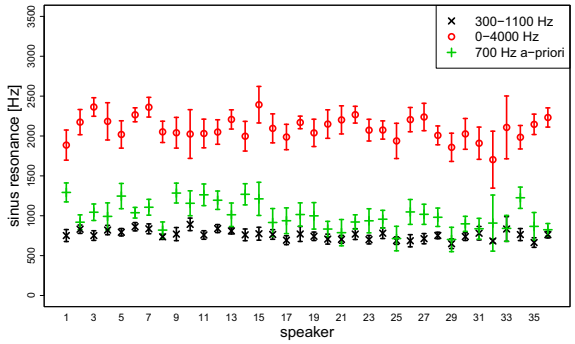

Figure 4: *Frequency estimates for different prior settings for the resonance frequency and for a low prior bandwidth and a small coupling area (compare upper left panel in Fig. 3). Errorbars denote 1.96 times the standard error.*

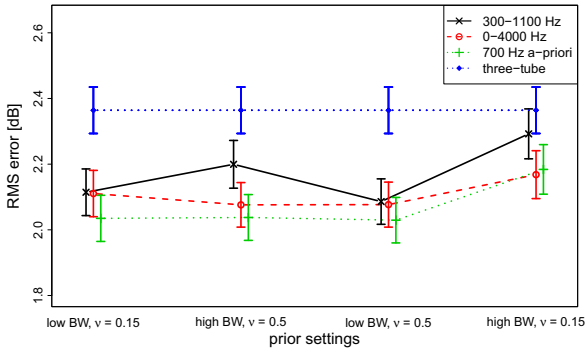

Figure 5: *RMS error as a function of prior settings. Errorbars denote 1.96 times the standard error and blue dots show the RMS error for the simple three-tube model.*

as it is 3-4 times higher than reported [5]. Third, the coupling area ratio ν for the sinus is either set a-priori to 0.5 (symmetric coupling) or to a significantly lower value of 0.15 (asymmetric coupling) which is closer to observed values [6]. Again, the sigmoidal transformation was used to restrict the estimates to their respective valid regions.

2.6. Evaluation data

To test the different a-priori settings nasal stops /n/ and /m/ were taken from 36 speakers of dialect region 1 from the TIMIT database [13]. This amounted to a total of 619 utterances (234 /m/ and 385 /n/). Speech segments of 40 ms around the center of the full nasal segment were taken. Segments with a duration of less than 40 ms were ignored. Signals were downsampled to 8000 Hz. A pre-emphasis was applied using an IIR filter with coefficients $(1, -0.9)$. The different model configurations as described above were estimated together with a simple three-tube model as shown in Fig. 2 without the maxillary sinus ($L = 4$, $M = 6$, and $N = 4$). The root mean square (RMS) of the estimation error in dB of the modeled transfer function with respect to the envelope was compared between the simple model and the different a-priori assumptions. Further, the estimated resonances of the model sinus were also compared.

3. Results

As expected, the estimation shows, that the a-priori assumptions do have a large influence on the resonance estimates of the modeled sinus (Figs. 3 and 4). In particular, the assumptions about the frequency range and the bandwidth affect the estimates of the paranasal cavity. An a-priori high bandwidth results in a much larger variation and higher values in the estimated bandwidth (Fig. 3). The effect of the assumptions about the frequency of the sinus is quite different. For the unrestricted frequency range the estimated frequencies lie much higher than what is to be expected for a maxillary sinus (red 'o' in both figures) even showing a slight bias towards frequencies above the prior value, i.e. 2 kHz. Setting the prior to a more realistic frequency (700 Hz) leads to a strong decrease of the estimated frequency (green '+'). The within-subject variance seems to be increased in both cases as compared to the restricted estimation, however, this is probably mostly due to the higher allowed range of frequencies. The effect of assumptions about coupling area ratio and bandwidth on the resonance frequency estimate is small compared to the effect of the frequency priors.

Comparing the simple three-tube model with the model including the sinus cavity shows that the RMS error is, on average, about 0.2 to 0.3 dB lower (Fig. 5) except for the case when the prior bandwidth is set high (i.e. the zero is a-priori assumed not to have a large influence on the transfer function) and the resonance frequency of the sinus is restricted to lie between 300 and 1100 Hz where the decrease in estimation error is less than 0.2 dB. Summarizing, the results show that, except for a high prior bandwidth, the prior assumptions have only a weak effect on the model error.

4. Discussion

Multi-branch models are powerful tools that allow to model a wide range of effects in the production of nasal speech. Their estimation, however, proves to be difficult simply due to the many degrees of freedom inherent in such models. Here, the effect of different prior assumptions for a paranasal cavity in a model for nasals was investigated with respect to the modeling error and the estimates of the resonance frequency and bandwidth of the paranasal cavity. This was done using a probabilistic model inversion scheme introduced previously.

The comparison of the model including one sinus compared to the three-tube model yielded a decrease in the overall estimation error that was only weakly dependent on the prior assumptions. From the results, however, it is clear that prior assumptions have a large influence on the estimated acoustic parameters of the paranasal cavity. In particular assumptions about the bandwidth and/or the resonance frequency yield quite different estimates of the these two quantities. In contrast, different assumptions about the coupling of the sinus that affects the resonances of the vocal tract model have, on average, only a comparatively small effect on the estimated acoustic parameters of the sinus.

Concluding, this study highlights the importance of adequate prior information in the estimation process of speech production models of nasal speech. A way to partially circumvent this problem may be to utilize the increased amount of data of multiple utterances via a joint estimation scheme. Preliminary data indicate a potential benefit, however, this approach needs to be investigated in more detail.

5. References

[1] T. Pruthi, C. Y. Espy-Wilson, and B. H. Story, "Simulation and analysis of nasalized vowels based on magnetic resonance imaging data." *The Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3858–3873, Jun. 2007.

[2] K. Schnell, "Rohrmodelle des Sprechtraktes. Analyse, Parameterschtzung und Syntheseexperimente," Ph.D. dissertation, Goethe University Frankfurt am Main, 2003.

[3] C. Kasess, W. Kreuzer, E. Enzinger, and N. Kerschhofer-Puhalo, "Estimation of the vocal tract shape of nasals using a bayesian scheme," in *Proceedings of Interspeech*, 2012.

[4] I.-T. Lim and B. Lee, "Lossy pole-zero modeling for speech signals," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 2, 1996.

[5] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation." *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3374–83, Nov. 1996.

[6] B. Story, "Physiologically-Based Speech Simulation Using An Enhanced Wave-Reflection Model of the Vocal Tract." Ph.D. dissertation, University of Iowa, 1995.

[7] E. Tarhan, M. Coskun, O. Cakmak, H. Celik, and M. Cankurtaran, "Acoustic rhinometry in humans: accuracy of nasal passage area estimates, and ability to quantify paranasal sinus volume and ostium size." *Journal of applied physiology*, vol. 99, no. 2, pp. 616–23, Aug. 2005.

[8] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-21, no. 5, pp. 417–427, 1972.

[9] J. Markel and A. Gray, Jr., *Linear Prediction of Speech*. Berlin: Springer, 1976.

[10] D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 237–248, 2010.

[11] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of Interspeech*, 2011.

[12] S. Julier and J. Uhlmann, "New extension of the Kalman filter to nonlinear systems," *Proceedings of the 1997 SPIE Conference on Signal Processing, Sensor Fusion, and Target Recognition*, vol. 3068, pp. 182–193, 1997.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Philadelphia: Linguistic Data Consortium, 1993.