LPC-BASED INVERSION OF THE DRM ARTICULATORY MODEL

Sacha KRSTULOVIĆ

IDIAP, C.P.592, CH-1920 Martigny, Switzerland
sacha@idiap.ch

ABSTRACT

Articulatory representations are expected to bring better speech recognition results. This requires to estimate the parameters of a speech production model from the speech sound, problem known as acoustico-articulatory inversion. Known methods to solve this problem usually introduce a heavy computational cost. Alternately, it is known that Linear Prediction analysis offers an analogy with acoustic filtering. This analogy had been exploited to develop a less expensive analytic method applicable to the estimation of tube shapes discretized in equal-length sections. We have extended the method to the DRM case, where the tube is made of unequal-length sections. The proposed DRM inversion scheme is thus simpler and faster. Furthermore, it shows good performance in terms of low residual modeling error. It also enhances speech recognition results when used to compute Log Area Ratios.

1. INTRODUCTION

It is traditionally recognized [3] that the production of speech results from the acoustical filtering of a glottal excitation by the vocal tract taken as a series of connected sections of uniform length. In this framework, it can be demonstrated [5] that the filtering process is equivalent to an Auto-Regressive (AR, or all pole) filtering process. The auto-regressive filter can be described in a polynomial form as well as in a lattice form. Wakita [11] has proposed to recover the tube shapes by applying inverse filtering methods, known in the framework of AR process identification, to acoustico-articulatory inversion.

Mrayati, Carré and Guérin [7] have pushed the above speech production theory further: they postulate that speech production modeling agrees better with human phonology's rules if the vocal tract is discretized in unequal-length sections. We will show in the following that the Distinctive Regions Model (DRM) they propose corresponds to a constrained AR process if the DRM tube is considered as a series of subsections of unit length, some of which being fastened together (Figure 1). First, the constrained lattice form of the DRM inverse transfer function is derived. A criterion to estimate its parameters is posed, and the corresponding estimator is expressed. Experimental assessment of the method is performed first through average mean square residual error computation and then through the use of DRM-derived features for speech recognition. Finally, the role of articulatory features in speech recognition is briefly discussed.

This work is supported by the Swiss FNRS grant Nr. 2100-49'725.96. The author is grateful to Dr. Chafic Mokbel for support to the present work.

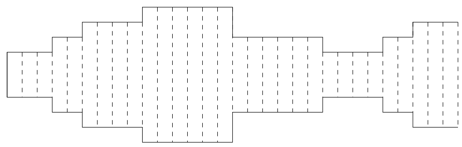

Figure 1: The DRM acoustic tube as a concatenation of 30 equal-length sections.

2. LATTICE FORM OF THE DRM FILTER

2.1. Development of the transfer function

The solution of Webster's equations, describing the state of a fluid in an excited acoustic tube [6, 1], allows to express the interaction between a sound wave propagating towards the lips end of a vocal tract and a sound wave propagating in the inverse direction, i.e. towards the glottal end. If the vocal tract is discretized in unequal length cylindrical section, this interaction is described by :

$$\begin{cases} u^+_{m+1}(t-\Delta_m t) = \frac{1}{1-k_m}\left\{u^+_m(t) + k_m u^-_m(t)\right\} \\ u^-_{m+1}(t+\Delta_m t) = \frac{1}{1-k_m}\left\{k_m u^+_m(t) + u^-_m(t)\right\} \end{cases} \quad (1)$$

where :

- u^+_m (resp. u^-_m) is the speed of the forward (resp. backward) traveling wave
- $k_m = \frac{S_{m+1}-S_m}{S_{m+1}+S_m}$, where S_m is the area of the m-th cylindrical section
- $\Delta_m t = \frac{\Delta l_m}{c}$ is the time needed for the air to travel along the m-th section.

By defining a discrete time unit corresponding to the greatest common divisor of the traveling times, the Z-transform can be applied to the above equations. As speed of sound is constant, the discrete time can be related to a discrete distance : $\Delta_{unit} t = \Delta l_{unit}/c$. This unit distance is itself the greatest common divisor of the section lengths. In the case of the DRM, this distance is $L/30$ (where L is the total tube length) since the sections have length $L/10$, $L/15$, $2L/15$, $L/5$ etc., hence the 30 unit sections in Figure 1. After application of the Z-transform, we obtain :

$$\begin{cases} z^{\frac{-n_m}{2}} U^+_{m+1}(z) = \frac{1}{1-k_m}\left[U^+_m(z) + k_m U^-_m(z)\right] \\ z^{\frac{n_m}{2}} U^-_{m+1}(z) = \frac{1}{1-k_m}\left[k_m U^+_m(z) + U^-_m(z)\right] \end{cases} \tag{2}$$

where :

- $z = e^{j\omega 2\Delta_{unit} t}$
- n_m is the length of the m-th section expressed as the number of elementary units.

This system can be expressed in matrix form as :

$$\begin{bmatrix} U^+_{m+1}(z) \\ U^-_{m+1}(z) \end{bmatrix} = \frac{z^{\frac{n_m}{2}}}{1-k_m} \begin{bmatrix} 1 & k_m \\ k_m z^{-n_m} & z^{-n_m} \end{bmatrix} \begin{bmatrix} U^+_m(z) \\ U^-_m(z) \end{bmatrix} \tag{3}$$

Considering that the exit of the tube (the "lips") is connected to a tube of infinite section, we obtain the following boundary condition[1] :

$$S_{-1} = \infty \;\Rightarrow\; k_0 = -1$$

Applying this condition, equation (3) can be written as :

$$\begin{bmatrix} U^+_{m+1}(z) \\ U^-_{m+1}(z) \end{bmatrix} = z^{\left(\frac{1}{2}\sum_{k=0}^{m} n_k\right)} K_m \begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix} \left\{U^+_0(z) - U^-_0(z)\right\} \tag{4}$$

with

$$\begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix} = \begin{bmatrix} 1 & k_m \\ k_m z^{-n_m} & z^{-n_m} \end{bmatrix} \begin{bmatrix} 1 & k_{m-1} \\ k_{m-1} z^{-n_{m-1}} & z^{-n_{m-1}} \end{bmatrix} \cdots \begin{bmatrix} 1 & k_1 \\ k_1 z^{-n_1} & z^{-n_1} \end{bmatrix} \begin{bmatrix} 1 \\ -z^{-n_0} \end{bmatrix} \tag{5}$$

and

$$K_m = \prod_{i=0}^{m} \frac{1}{1-k_i} \tag{6}$$

Neglecting the overall delay $z^{\left(\frac{1}{2}\sum_{k=0}^{m} n_k\right)}$ and the gain K_m, *the true transfer function for the forward traveling wave corresponds to* $A(z) = 1/D^+_m(z)$, *where* $D^+_m(z)$ *is computed recursively using* :

$$\begin{aligned} \begin{bmatrix} D^+_{m+1}(z) \\ D^-_{m+1}(z) \end{bmatrix} &= \begin{bmatrix} 1 & k_{m+1} \\ k_{m+1} z^{-n_{m+1}} & z^{-n_{m+1}} \end{bmatrix} \begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix} \\ \begin{bmatrix} D^+_0(z) \\ D^-_0(z) \end{bmatrix} &= \begin{bmatrix} 1 \\ -z^{-n_0} \end{bmatrix} \end{aligned} \tag{7}$$

It can be demonstrated [4] that this recursion leads to a polynomial form for $D^+_m(z)$. Hence, the "synthesis oriented" transfer function $A(z)$ still represents an autoregressive process. Furthermore, the inverse transfer function $D^+_m(z)$ can be formalized as a lattice, given in Figure 2.

[1] Sections are numbered in crescent order from lips to glottis.

The given lattice form is equivalent to a classical inverse Linear Prediction lattice filter where some reflection coefficients k_i would be constrained to stay equal to zero in the locations corresponding to fastened unit sections. It has been demonstrated [5] that unconstrained AR filters where stable if the condition $|k_i| < 1 \; \forall i$ was verified. Since the constraint introduced by the DRM means $|k_i| = 0$ for some i, the stability of the model is still guaranteed if the unconstrained reflection coefficients are between 1 and -1.

Since we have now a lattice form and a stability condition for the transfer function, we can apply inverse filtering to the estimation of its parameters.

3. INVERSE FILTERING INCLUDING DRM CONSTRAINTS

We have just shown that the DRM acoustic filtering process is equivalent to a lattice filtering process including articulatory constraints in the form of odd delays (Fig. 2). Denoting by Σ_p the sum of all the delays from order 1 to order (p), a stable filter characterized by its reflection coefficients $k_{(p+1)}$ can be estimated by minimizing a mean squared prediction error $\xi(p+1)$ similar to the one used for Burg's method :

$$\xi^2(p+1) = \frac{1}{2}\left\{\sum_{t=\Sigma_p+1}^{N} \epsilon^+_t(p+1)^2 + \sum_{t=\Sigma_p+1}^{N} \epsilon^-_t(p+1)^2\right\} \tag{8}$$

Minimizing this error criterion through differentiating and equating to zero gives :

$$k_{p+1} = \frac{-2\sum_{t=\Sigma_p+1}^{N} \epsilon^+_t(p)\epsilon^-_{t-n_p}(p)}{\sum_{t=\Sigma_p+1}^{N}\left(\epsilon^+_t(p)\right)^2 + \sum_{t=\Sigma_p+1}^{N}\left(\epsilon^-_{t-n_p}(p)\right)^2} \tag{9}$$

It can be easily verified that this solution always respects the filter stability condition evoked in the preceding section.

4. EXPERIMENTAL RESULTS

4.1. Data set

The described inverse filtering method has been implemented with the purpose of comparing its modeling accuracy with other models. Comparison has been performed between a classical 8th order LP lattice (LPC8), a 27th order LP lattice (LPC27), the DRM lattice and an evenly-constrained lattice (all z^{-n_i} equal to z^{-4}; denoted LPC Constr.).

For all our experiments, a subset of the sound files of the University of Wisconsin X-Ray microbeam database [12] has been used. Characteristics of the data are : microphone speech w/ some background noise, 47 speakers

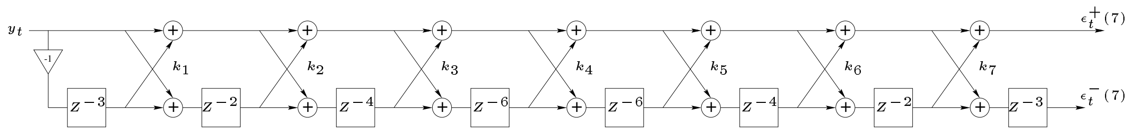

Figure 2: The DRM inverse filter.

(21 M, 26 F), 21.239 kHz sampling rate, prompted word lists (7 words or less). This choice has been operated in conjunction with other tasks of our research project.

4.2. Summary of the analysis method

The speech analysis method corresponding to the compared models can be summarized as :

1. *adapt the sample frequency of speech data* to the constraint imposed by the model for the application of the Z-transform, namely $F_s = \frac{c}{2\,\Delta l_{unit}}$ (where c is the speed of sound and Δl_{unit} is the unit length defined above). Given the structure of the models, and assuming that a vocal tract is 17 centimeters long on average, the sampling frequencies to use will be 8kHz in the case of LPC8 and 30kHz in the cases of the DRM, LPC27 and LPC Constr. The frequency adaptation has been performed from the unaltered original 21kHz sampling rate, using zero packing followed by polyphase filtering.

2. *pre-emphasize the obtained speech wave* by a simple differentiation. This step is performed to compensate for the effects introduced by the glottal waveform shape and the radiation effect at the lips.

3. *inverse-filter speech*. Adapt filter every 10 ms by computing reflection coefficients k_i, using expression (9) with 25 ms observation windows. For speech recognition, transform reflection coefficients into log area ratios : $l_i = log(\frac{1-k_i}{1+k_i})$.

4. *compute average mean squared prediction error* in the different cases studied.

4.3. Residual prediction error

Results (Figure 3) show that the DRM, described by 7 parameters (the k_i's), produces a lower prediction error than an unconstrained LPC model described by the same number of parameters, i.e. LPC8.

However, an unconstrained LPC model of an equivalent order (LPC27) results, as could be expected, in a lower residual error.

Nevertheless, it is seen that the repartition of the constraints in the section lengths plays a role in the modeling accuracy, since a constrained model of equivalent order but with a different repartition of section lengths (LPC Constr.) performs worse than the DRM.

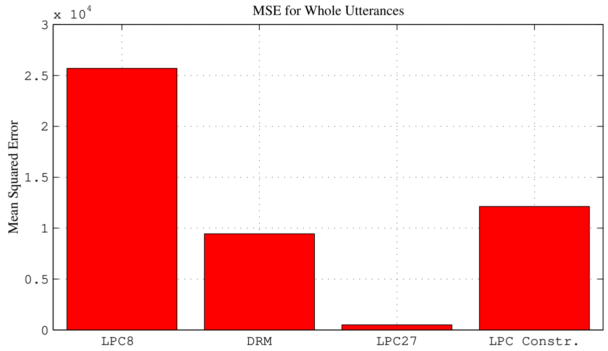

Figure 3: Comparison of Mean Squared Errors for the different tube models. Upsampling or downsampling is performed from the original 21 kHz sample frequency.

4.4. Speech recognition

Speaker-independent speech recognition has been performed by training 43 monophone HMMs (left-right, 3 states, 5 gaussians per state, diagonal covariances) in an embedded training scheme, using the orthographic transcription given with the database. Training set was made of 1584 sentences divided among 41 speakers (18 M, 23 F). Test set was made of 237 sentences/6 speakers. Decoding concerned the 110 words vocabulary, and made use of a descriptive grammar giving the number of words (7, 6 or 5). Our baseline system, making use of LPCC+Δ+$\Delta\Delta$ features, reached a 5.8% Word Error Rate (WER).

Log Area Ratios derived from the DRM have been used for recognition and compared to LAR from an 8th order equal-length tube. While unconstrained LAR gave a 16.2% WER, DRM-LAR gave a 13.4% WER, *which corresponds to a 17% relative improvement*.

Of course, these results appear low as compared to the baseline system. However, they suggest that if a reliable way of computing cepstral coefficients from the DRM parameters is found, an improvement in state-of-the-art recognition accuracy could be expected.

5. DISCUSSION

Many authors [2, 8, 13] have suggested that speech recognition would benefit from the exploitation of knowledge related to the capabilities of the speech production apparatus. As the direct observation of vocal tract shapes is impractical for everyday life applications, the observation of articulation through acoustic clues, problem known as acoustico-articulatory inversion, is contemplated as a means of obtaining “speech production clues”.

Solutions to this problem had already been proposed [10] through the use of optimization, artificial neural networks or codebook based methods. Richards & al. [9] have used a dynamic codebook search to recover the parameters of the DRM. These methods introduce a heavy computational cost. Alternately, Wakita [11] had developed a less expensive analytic method applicable to the estimation of tube shapes discretized in equal-length sections.

What we have proposed here, through the extension of Wakita's method to the DRM case, is a fast and simple DRM inversion scheme. But in the absence of a definition for an "articulatory distance", we wish to keep on using acoustic features at the input of recognition systems. Hence, we believe that speech recognition won't benefit from the direct use of articulatory parameters as patterns to be recognized, but rather from imposing articulatory constraints on the observation of acoustic features.

6. CONCLUSION

We have proposed a simple and fast method to invert the DRM speech production model. Performance of the inversion system is good in terms of low mean squared residual error. Furthermore, an improvement is observed in speech recognition accuracy when using DRM-derived Log Area Ratios in place of usual LAR. This method can be useful in several speech processing domains, including speech coding and speech recognition, and preliminary experiments validate the approach. But significant improvements in speech recognition still depend upon the finding of a reliable way to compute cepstral coefficients from DRM features.

References

[1] L.J. Bonder. The n-tube formula and some of its consequences. *Acustica*, 52:216–226, 1983.

[2] K. Erler and L. Deng. HMM representation of quantized articulatory features for recognition of highly confusible words. In *ICASSP '92*, volume I, pages 545–548, 1992.

[3] C.G.M. Fant. *Acoustic theory of speech production*. Mouton: The Hague, 1960.

[4] Sacha Krstulović. Acoustico-articulatory inversion of the DRM model through inverse filtering. IDIAP-RR 16, IDIAP, 1998. `http://www.idiap.ch/`.

[5] J.D. Markel and A.H. Gray. *Linear prediction of speech*. Springer-Verlag, 1976.

[6] P.M. Morse and K.U. Ingard. *Theoretical acoustics*. Mc Graw-Hill, 1968.

[7] M. Mrayati, R. Carré, and B. Guérin. Distinctive regions and modes: a new theory of speech production. *Speech Communication*, (7):257–286, 1988.

[8] G. Papcun, J. Hochberg, T.R. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *Journal of the acoustical Society of America*, 92:688–700, 1992.

[9] H. B. Richards, J. S. Mason, M. J. Hunt, and J. S. Bridle. Deriving articulatory representations of speech with various excitation modes. In *ICSLP'96*, volume 2, pages 1233–1236, 1996.

[10] J. Schroeter and M.M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1, part II):133–150, 1994.

[11] H. Wakita. Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(3):281–285, 1979.

[12] J.H. Westbury. *X-ray microbeam speech production database user's handbook*. Waisman Center, University of Wisconsin, 1.0 edition, June 1994.

[13] I. Zlokarnik. Experiments with an articulatory speech recognizer. In *Eurospeech '93*, pages 2215–2218, 1993.