Literature Review

Yonatan Vaizman

I. Audio processing

[1] for speech coding tutorial.....

[2] for joint spectral-envelope and f0 estimation....

[3] efficient solution to sparse linear prediction analysis for speech...

Books: [4]–[7]

A. LPC

Good tutorial for linear prediction framework [8]. [9], [10]... [11], [12] sparse linear prediction for speech....

B. Vocoders

[13]....

C. Frequency-warped LPC

D. Fourier space excitation-filter framework

E. Filterbanks (overcomplete codebooks?) for speech analysis

Speech coding with VQ [10]....

II. Optimization and learning

A. LPC optimization

Good tutorial for linear prediction framework [8]. Makhoul talks about the squared error criterion, its advantages and shortcomings and when it might not fit the best spectral envelop (when the excitation input to the system is a pitched periodic pulse train — in which case the peaks of the estimated filter coincide with the harmonics instead of the better fitting envelop).

In [14] El-Jaroudi *et al.*give intuition to the inappropriateness of traditional linear prediction with L2 minimization. They discuss the case of signals that have discrete spectra, such as voiced speech, which theoretically only has power in the fundamental frequency and in its harmonic multiples. They regard to more general cases, where there is only power in a discrete, finite set of frequencies (not necessarily integer multiples of a fundamental). According to the excitation-filter model such observed signals were generated when the resonating all-pole filter was excited with a discrete-spectrum excitation signal. Since the power of the system is only sampled at discrite points, the autocorrelation of the observed signal has strong aliasing and repetitions. This aliasing is more severe when the frequency-sampling is more sever (when the fundamental is a higher frequency). Because of this aliased version of the original system's autocorrelation (inverse transform of the filter's power spectrum), the unique solution of filter coefficients that the LP finds is not the original ("true") filter that generated the signal. In their work the authors suggested Discrete All-Pole (DAP) modeling, in which an autocorrelation signal is calculated only using the specific frequencies that have effective power (as a discrete inverse transform — sum over the finite set of "on" frequencies, instead of integral over all frequencies). They optimize the Itakura-Saito error measure in its discrete-frequency version, and suggest an iterative algorithm to solve for the optimal filter.

However, the usage of their algorithm requires to first estimate the power spectrum of the observed signal and perform peak picking to locate the discrete (and finite) set of frequencies to be used in the algorithm. In practice, when noisy observations are present, the stage of peak picking depends on other algorithms and may introduce biases and errors.

B. L2, L1, Lp minimizations

In [15] Denoel and Solvay analyze using linear prediction with L1 minimization (least absolute error criterion) for speech. This optimization problem is still convex, but usage of Linear Programming to solve it requires heavy computation and doesn't guarantee that the selected optimal filter is stable. The authors use a lattice structure and derive a Burg-like order-recursive algorithm to optimize the reflection coefficients one after the other. This suggested algorithm does guarantee a stable filter for every order. The heavy computation of this algorithm lies in a median calculation for every order. When the application is coding for transmission, this can be alleviated by replacing sorting of values with bisection to bins (according to the number of bits provided for the coding of a reflection coefficient), and the leftover error can be compensated in the next order. However, if quantization and coding are not done and simply the optimal filter is estimated, this method is still computationally heavier than L2 methods.

Linear prediction with L1 norm [16]...

Adaptive Lp linear prediction [17]...

[11], [12] sparse linear prediction for speech....

Stable IIR design based on Lp error minimization [18]....

Giri and Rao - block sparse excitation criterion [19]...

C. Optimization of mixtures

D. Sparsecoding and compressed sensing

E. LTI filter clustering

Perceptually consistent measures of spectral distance [20]....

Speech coding with VQ [10]....

Spectral distance measures [20]....

VQ in speech coding [21]...

III. MIR applications

A. Source separation

[22]....

[23] ....

B. Multiple pitch (fundamental frequency) estimation

In [24] the authors calculate a salience function for every period by summing spectral amplitudes with modeled weights (by training and fitting over examples they concluded a suitable model should be linearly increasing as function of the candidate fundamental frequency and decreasing like $\frac{1}{m}$ as a function of the harmonic m). After calculating the salience of different possible fundamental frequencies they propose three different methods to find the correct fundamental frequencies in the signal. First a direct method detects the maximal salience values. Second an iterative method detects the most likely fundamental frequency and then cancels its (weighted) contribution to the sound from the mixture and continues to find the next. Third, a joint estimation detects the f0 values together. As pre-processing they use spectral flattening.

C. Melody extraction and automatic transcription

[25]–[27].... Guitar chords and fingering [28]....

D. Chord recognition

E. Instrument recognition

Classification of pitched musical instruments [29]... [30], [31]....

Polyphonic and polyinstrument [32]....

Instrument recognition (temporal and cepstral features) [33].... and comparison of acoustic features [34]....

Identifying woodwind instruments [35].....

instrument recognition and affect on mir [36]....

temporal integration [37]....

Multitrack mixing: [38], [39]....

In [40] isolated sound single instrument recognition was performed using recordings from UIowa, McGill, OLPC and Freesound datasets, with total of 24 instruments. For each separate note (or beat) first STFT was calculated, then non-negative matrix factorization (NMF) was performed on the magnitude STFT. Each atom of the approximation factorization (each combination of spectral column and temporal row) was analyzed separately in the spectral component (the column vector was processed to produce 32 MFCCs) or in the temporal component (the row vector was processed with a predetermined 32-channel Gamma filterbank, and from each response signal the maximal value was retained, for time-shift invariance). RBF-kernel SVM classification was done either using only spectral features (32 MFCCs of an atom), only temporal features (32 Gamma response maxima of an atom), or concatenation of both (64 dimensions per atom). Each atom was regarded as an instance and classified. In their results they observed generally better performance by spectral features compared to temporal features, but significant improvement when using both types of features.

In [41] Fuhrmann *et al.*compared different methods to select segments out of a musical excerpt (full production), and apply predominant instrument recognition only on those selected segments. The methods were: taking all the excerpt (regarded as "upper bound"), taking the first 30 seconds, taking uniformly spaced segments and clustering segments. They shown that selecting with clustering is better than the other segment-selection methods. The whole point of using selected segments was efficiency: to analyse less audio time (taking advantage of the redundancy and repetitiveness common in songs). However, in order to "cleverly" select the segments additional analysis should be done (the clustering), using features from all the excerpt, and the selected segments in the clustering methods occupied, in average, 0.66 of the audio time, which is still a lot of time to analyze. In addition, the experiments were done only on 220 excerpts of 30 seconds each, and the results don't show much of an improvement compared to the runtime-cheaper segment-selection methods.

In [42] Bosch *et al.*suggested combining a pre-stage of an of-the-shelf source separation algorithm (FASST - Flexible Audio Source Separation Framework by Ozerov *et al.* [23]) before performing instrument recognition. They applied standard separation to 4 channels: bass, drums, melody and other. They observed that when using the same models for instrument recognition on each channel and then combining, the results are worse than using the original audio. However, when training SVM models separately on each of the 4 channels, the result is an improvement.

In [31] Yu *et al.*suggested encoding real-cepstrum coefficients using sparse coding, for instrument recognition. They also saw improvement when compressing the magnitude spectrum with a power law (with powers $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$) instead of logarithmic compression. The resulted cepstrum doesn't have the theoretical insight of linearly additive excitation and filter components (as the log-cepstrum has) but it still works well for instrument recognition.

IV. Datasets

References

[1] A. S. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[2] H. Kameoka, N. Ono, and S. Sagayama, "Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1507–1516, 2010.

[3] V. Khanagha and K. Daoudi, "An efficient solution to sparse linear prediction analysis of speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–9, 2013.

[4] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-hall Englewood Cliffs, 1978, vol. 100.

[5] W. C. Chu, *Speech coding algorithms: foundation and evolution of standardized coders*. John Wiley & Sons, 2004.

[6] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. Elsevier Science Inc., 1995.

[7] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River (NJ, USA): Prentice Hall, 1993.

[8] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[9] ——, "Stable and efficient lattice methods for linear prediction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 25, no. 5, pp. 423–428, 1977.

[10] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1551–1588, 1985.

[11] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing." in *INTERSPEECH*, 2008, pp. 1353–1356.

[12] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1644–1657, 2012.

[13] J. Makhoul, R. Viswanathan, and W. Russell, "A framework for the objective evaluation of vocoder speech quality," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76.*, vol. 1. IEEE, 1976, pp. 103–106.

[14] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *Signal Processing, IEEE Transactions on*, vol. 39, no. 2, pp. 411–423, 1991.

[15] E. Denoel and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 6, pp. 1397–1403, 1985.

[16] J. Schroeder and R. Yarlagadda, "Linear predictive spectral estimation via the l_1 norm," *Signal processing*, vol. 17, no. 1, pp. 19–29, 1989.

[17] J. Lansford and R. Yarlagadda, "Adaptive l_p approach to speech coding," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 335–338.

[18] C.-C. Tseng, "Design of stable iir digital filter based on least p-power error criterion," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 51, no. 9, pp. 1879–1888, 2004.

[19] R. Giri and B. D. Rao, "Block sparse excitation based all-pole modeling with applications to speech," in *ICAASP*, 2014.

[20] R. Viswanathan, J. Makhoul, and W. Russell, "Towards perceptually consistent measures of spectral distance," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76.*, vol. 1. IEEE, 1976, pp. 485–488.

[21] A. Gersho, S. Wang, and K. Zeger, "Vector quantization techniques in speech coding," *Advances in Speech Signal Processing, S. Furui and MM Sondhi, eds., Mar-cel Dekker, New York*, pp. 49–84, 1992.

[22] M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 2. IEEE, 1994, pp. II–77.

[23] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1118–1133, 2012.

[24] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes." in *ISMIR*, 2006, pp. 216–221.

[25] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[26] K. O'Hanlon and M. D. Plumbley, "Automatic music transcription using row weighted decompositions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 16–20.

[27] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.

[28] A. M. Barbancho, A. Klapuri, L. J. Tardon, and I. Barbancho, "Automatic transcription of guitar chords and fingering from audio," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 915–921, 2012.

[29] P. Herrera-Boyer, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal processing methods for music transcription*. Springer, 2006, pp. 163–200.

[30] K. D. Martin and Y. E. Kim, "Musical instrument identification: A pattern-recognition approach," *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1768–1768, 1998.

[31] L.-F. Yu, L. Su, and Y.-H. Yang, "Sparse cepstral codes and power scale for instrument identification," in *Proc. ICASSP*, 2014.

[32] P. Hamel, S. Wood, and D. Eck, "Automatic identification of instrument classes in polyphonic and poly-instrument audio." in *ISMIR*. International Society for Music Information Retrieval conference (ISMIR), 2009, pp. 399–404.

[33] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 2. IEEE, 2000, pp. II753–II756.

[34] A. Eronen, "Comparison of features for musical instrument recognition," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001, pp. 19–22.

[35] J. C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1064–1072, 2001.

[36] T. Kitahara, "Computational musical instrument recognition and its application to content-based music information retrieval," *Unpublished PhD Thesis, Kyoto University, Kyoto, Japan. Retrieved*, vol. 10, no. 31, p. 07, 2007.

[37] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 174–186, 2009.

[38] J. Scott and Y. E. Kim, "Instrument identification informed multi-track mixing," in *International Society for Music Information Retrieval conference (ISMIR)*, 2013.

[39] J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim, "Automatic multi-track mixing using linear dynamical systems," in *Proceedings of the 8th Sound and Music Computing Conference, Padova, Italy*, 2011.

[40] S. K. Tjoa and K. R. Liu, "Musical instrument recognition using biologically inspired filtering of temporal dictionary atoms." in *ISMIR*, 2010, pp. 435–440.

[41] F. Fuhrmann and P. Herrera, "Quantifying the relevance of locally extracted information for musical instrument recognition from entire pieces of music." in *ISMIR*, 2011, pp. 239–244.

[42] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." in *ISMIR*, 2012, pp. 559–564.