Ψηφιακή Επεξεργασία Ήχου
Μάθημα 9:Διαχωρισμός Αρμονικών και κρουστικών στοιχείων

Π.Μ.Σ. «Τεχνολογίες Ήχου και Μουσικής»

Δρ. Χρυσούλα Αλεξανδράκη

Τμήμα Μουσικής Τεχνολογία και Ακουστικής

Ελληνικό Μεσογειακό Πανεπιστήμιο

6/12/2024

Cocktail Party Effect

- Cocktail Party Effect
  - Η αντιληπτική ικανότητα να διαχωρίζουμε έναν ήχο μέσα από ένα σύνολο ήχους και να προσηλωνόμαστε σε αυτόν
  - Έχει αποδοθεί σε νευρολογικά χαρακτηριστικά και είναι πιο αποτελεσματική στην περίπτωση της αμφίπλευρης ακοής (binaural hearing)
    - Άρα θεωρείται ότι εν μέρει οφείλεται στη δυνατότητα εντοπισμού ηχητικών πηγών στο χώρο
  - Μελετάται ερευνητικά από τη δεκαετία του ‘50

Audio Source Separation

- Ο διαχωρισμός ηχητικών σημάτων που προέρχονται από διαφορετικές πηγές από ένα συνολικό σήμα που έχει προέλθει από ηχογράφιση ή/και μίξη, με καθαρά υπολογιστικές μεθόδους
- Εφαρμογές
  - Σε σήμα ομιλίας
  - Σε μουσικό σήμα
  - Σε περιβαλλοντικούς ήχους
- Μαθηματική έκφραση προβλήματος:
  - Να βρεθούν σήματα x_1, $x_{2, \ldots,}$ έτσι ώστε από δεδομένο σήμα x,
    - $x = x_1 + x_2 + \ldots$

Προσεγγίσεις

- **Blind Source Separation**
  - Δεν προϋποθέτει προηγούμενη γνώση (ή μόνο πολύ μικρή γνώση) για το περιεχόμενο του σήματος
  - Εφαρμογές
    - Image Analysis, Communications, Stock Prediction, Seismic Monitoring
  - Μαθηματικές προσεγγίσεις
    - Non-Negative Matrix Factorization (NMF), Principal Component Analysis (PCA), κ.α.
- **Informed Source Separation**
  - Αξιοποιεί γνώση για το τι περιέχεται στο σήμα, προκειμένου να διαχωρίσει διαφορετικές πηγές
  - Παραδείγματα
    - Score informed Source Separation (music), Text Informed Source Separation (speech), User Guided Source Separation (π.χ. θέλω να εντοπίσω την τάδε ηχητική πηγή) κ.α.

Harmonic Percussive Source Separation (HPSS)

- Αποσύνθεση Ήχου (Audio Decomposition)
  - $x = x_h + x_p$, όπου:
    - x_h : σήμα που περιέχει τα αρμονικά στοιχεία του αρχικού σήματος -> υψηλή κατά τόπους περιοδικότητα
    - x_p : σήμα που περιέχει μη αρμονικά στοιχεία -> υψηλή περιεκτικότητα σε υψηλές συχνότητες μικρής διάρκειας
- Επιτελείται σε μουσικό σήμα, όχι τόσο για το διαχωρισμό των μουσικών οργάνων, όσο ως:
  - Pre-processing step για άλλα tasks, π.χ.
    - Onset detection, Audio Transcription, Instrument Identification, κ.λπ.

Φασματογράφημα -> Περιοδικότητα και χρονική ακρίβεια

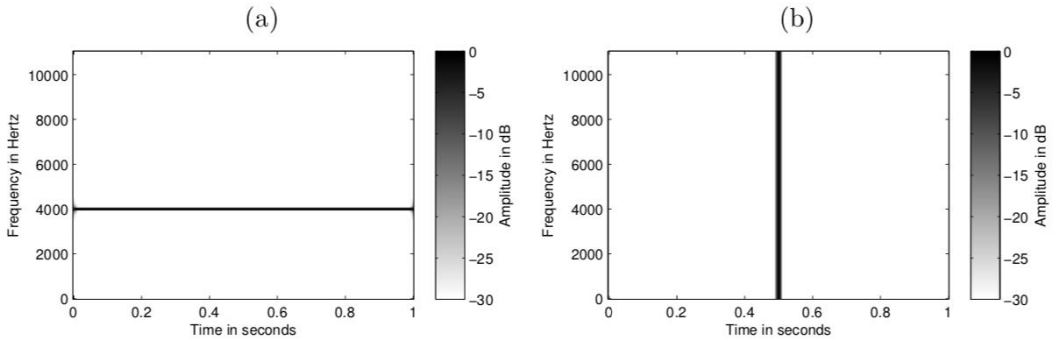


Figure 1: **(a):** Spectrogram of an ideal harmonic signal. **(b):** Spectrogram of an ideal percussive signal.

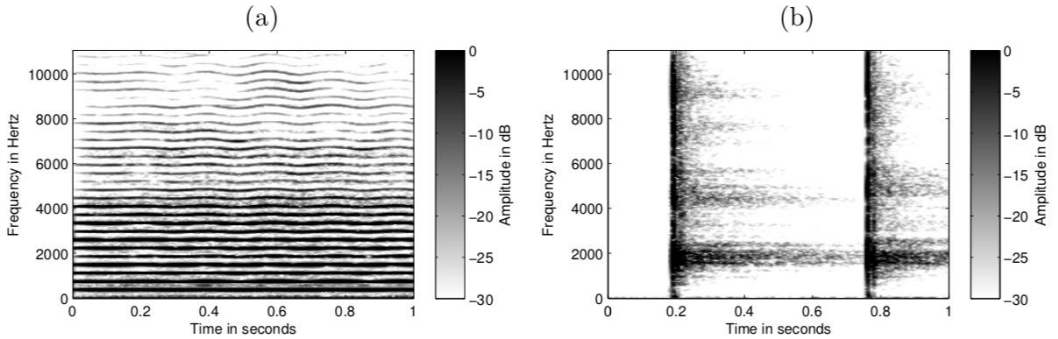
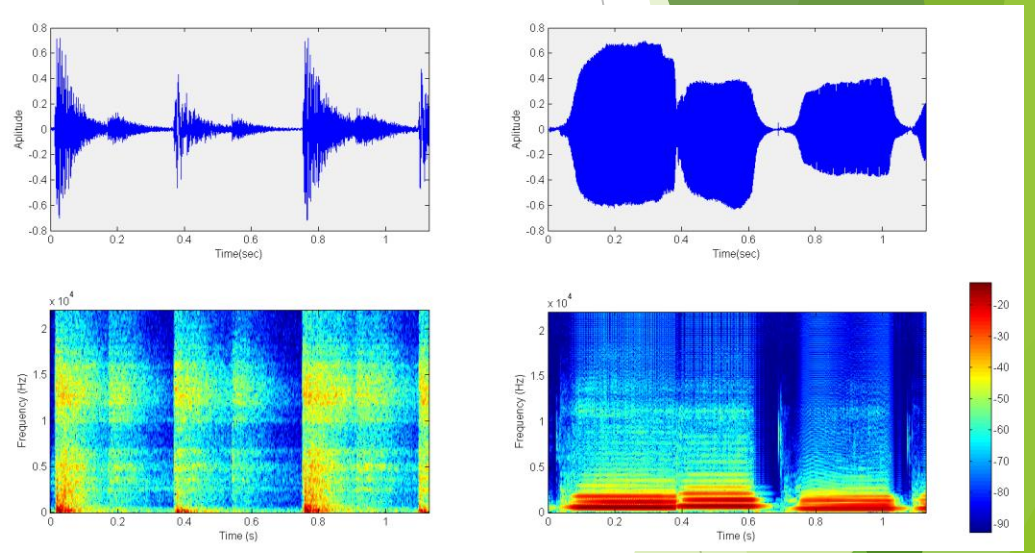


Figure 2: **(a):** Spectrogram of a recording of a violin. **(b):** Spectrogram of a recording of a castanets.



Drum Signal=> Salient Onsets
Flute Signal => Salient Pitches

Block Diagram HPSS

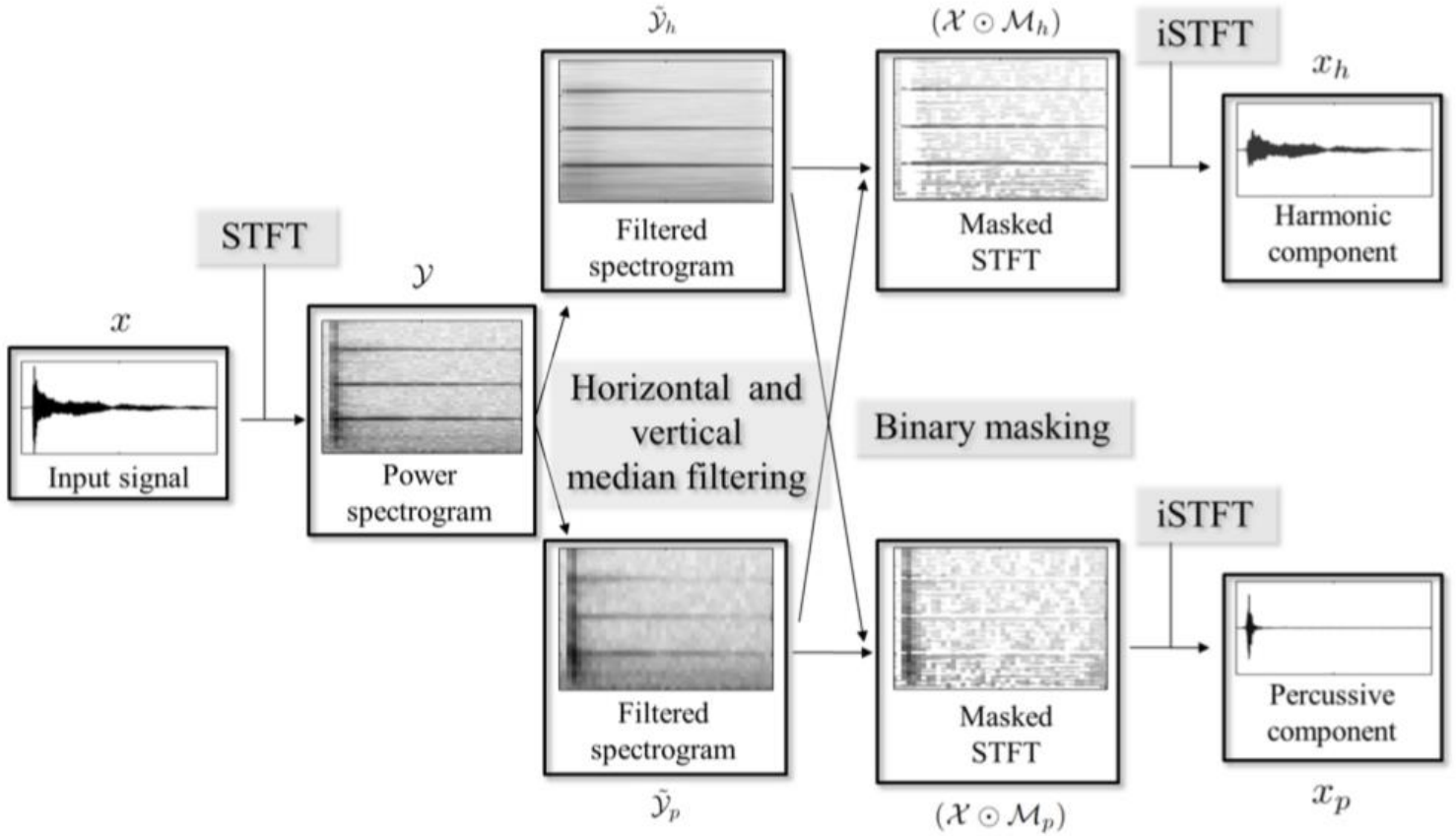


Figure 3: Harmonic-percussive source separation.

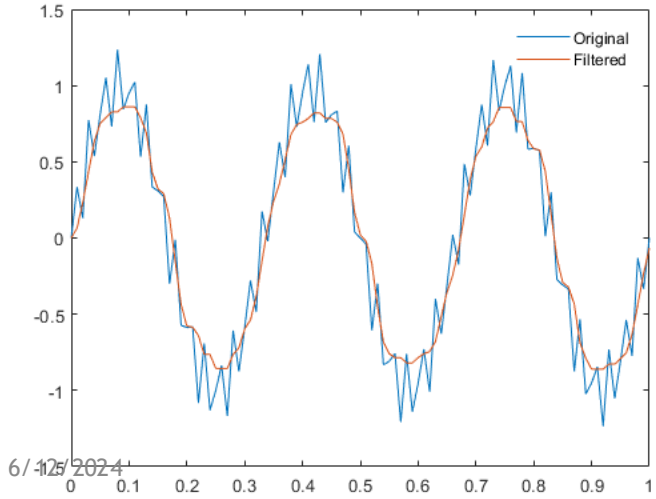
Müller, M. Lab course 2014

Αλγόριθμος HPSS

1. **STFT** στο αρχικό σήμα => X(k, n) Μιγαδικός πίνακας δύο διαστάσεων: A/(N-h) x N/2 +1 που οπτικοποιείται με το φασματογράφημα
2. Υπολογίζω το **power spectrogram** ως:
   - **Y(k, n) =** $|X(k,n)|^2$
3. Ο προκύπτον πίνακας φιλτράρεται με ένα **smoothing filter** σε δύο διαστάσεις:
   - Smoothing (Λείανση) στην οριζόντια διάσταση (horizontal smoothing) διατηρεί τα κυρίαρχα στοιχεία στη συχνότητα => Harmonic Spectrogram $Y_h(k, n)$
   - Smoothing στην κάθετη διάσταση (vertical smoothing) διατηρεί τα κυρίαρχα στοιχεία στο χρόνο => Percussive Spectrogram $Y_p(k, n)$
4. Εφαρμόζω την τεχνική του **Masking** ώστε να προκύψουν δύο πίνακες:
   - Ο harmonic mask $M_h(k, n)$ ο οποίος υποδεικνύει τα σημεία πίνακα στα οποία $Y_h(k, n) >= Y_p(k, n)$
   - Ο percussive mask $M_p(k, n)$ ο οποίος υποδεικνύει τα σημεία πίνακα στα οποία $Y_p(k, n) > Y_h(k, n)$
5. Εφαρμόζω **Inverse STFT** στον πολ/σμο πινάκων του X(k, n) με τους δύο masks για να πάρω δύο σήματα:
   - x_h = ISTFT $\{M_h(k, n)^* X(k, n)\}$
   - x_p = ISTFT $\{M_p(k, n)^* X(k, n)\}$

Smoothing – Median Filtering

- Για smoothing filter χρησιμοποιούμε αυτό που ονομάζεται **Median Filter**, το οποίο έχει ως στόχο να αφαιρέσει θόρυβο από ένα σήμα.
  - Έχει χρησιμοποιηθεί εκτενώς σε επεξεργασία εικόνας καθώς έχει δύο πλεονεκτήματα:
    1. Removes Noise
    2. Preserves Edges
- Βασίζεται στην έννοια του Αριθμητικού Μέσου (ΑΜ) median ενός συνόλου αριθμών
- Για καλύτερη ίσως επεξήγηση διαβάστε το άρθρο
  - What is the math behind median filter's noise reduction property?



Median – Διάμεσος

- Ορισμός:
  - Για ένα σύνολο αριθμών, ο **ΔΙΑΜΕΣΟΣ** είναι εκείνος ο αριθμός για τον οποίο οι μισοί αριθμοί του συνόλου είναι μικρότεροι από αυτόν ενώ οι άλλοι μισοί μεγαλύτεροι από αυτόν
- Μαθηματική Έκφραση:
  - Εάν $A = \{a_n \in \mathbb{R} | n \in [0, N-1]\}$ σύνολο αριθμών που έχουν ταξινομηθεί σε αύξουσα σειρά, δηλ. a_n <= $a_{n'}$ όταν n<n', τότε ο ΑΜ αυτού είναι ο αριθμός:
  - $$median(\mathrm{A}) := \begin{cases} a_{\frac{\mathrm{N}-1}{2}}, \ για \ \mathrm{N} \ περιττό \\ \frac{1}{2}(a_{\frac{\mathrm{N}}{2}} + a_{\frac{\mathrm{N}}{2}+1}), για \ \mathrm{N} \ άρτιο \end{cases}$$

Median vs. Mean (Average)

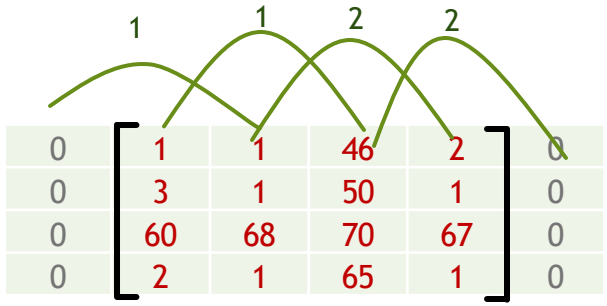
- Arithmetic Mean (Διάμεσος)
  - Στη στατιστική αποτελεί μία ένδειξη για το διάστημα στο οποίο κινούνται οι αριθμοί σε ένα διάστημα (υποσύνολό τους), αγνοώντας ακραίες τιμές (outliers)
- Median or Average (Μέσος όρος ή απλά μέσος)
  - λαμβάνει υπόψη εξίσου όλες τις τιμές ενός συνόλου αριθμών
- Παραδείγματα: να υπολογιστεί ο mean και ο median:
  - A = {2, 3, 190, 2, 3}, B = {3, 5, 78, 6}
  - Εφαρμόστε τον οριζόντιο (ή τον κάθετο) αριθμητικό μέσο μήκους 3 στον πίνακα
  - Hint: για κάθε στοιχείο του πίνακα πρέπει να βρείτε 3 γειτνιάζοντα στοιχεία, αν δεν υπάρχουν θεωρήστε ότι είναι μηδενικά

$$B = \begin{bmatrix} 1 & 1 & 46 & 2 \\ 3 & 1 & 50 & 1 \\ 60 & 68 & 70 & 67 \\ 2 & 1 & 65 & 1 \end{bmatrix}$$

Παράδειγμα – Υπολογισμός Διάμεσου σε πίνακα 2D

Να εφαρμοσθεί στον πίνακα Β, φίλτρο αριθμητικού μέσου διάστασης 1x3

$$B = \begin{bmatrix} 1 & 1 & 46 & 2 \\ 3 & 1 & 50 & 1 \\ 60 & 68 & 70 & 67 \\ 2 & 1 & 65 & 1 \end{bmatrix}$$



1	1	2	2

Ερμηνεία Horizontal/Vertical Smoothing

- Στον πίνακα αυτό που προκύπτει από STFT, διαφορετικές **γραμμές** αντιστοιχούν σε διαφορετικές ζώνες συχνοτήτων (frequency bins), ενώ διαφορετικές **στήλες** αντιστοιχούν σε διαφορετικά audio blocks.

$$Y_{kn} = \begin{bmatrix} Y_{1,1} & \cdots & Y_{1,A/(N-h)} \\ \vdots & \ddots & \vdots \\ Y_{\frac{N}{2}+1,1} & \cdots & Y_{\frac{N}{2}+1,,A/(N-h)} \end{bmatrix}$$

Τιμές ενέργειας για μία συχνότητα, ως προς το χρόνο

Τιμές ενέργειας για μία χρονική στιγμή, ως προς τη συχνότητα

- Οριζόντια (χρονική) λείανση -> κρατάει τις τιμές ενέργειας στις συχνότητες που κυριαρχούν σε ικανά χρονικά διαστήματα => αρμονικά στοιχεία $Y_h(k, n)$
- Κάθετη (συχνοτική) λείανση -> κρατάει τις τιμές ενέργειας στις χρονικές στιγμές που κυριαρχούν για ικανό εύρος συχνοτήτων => κρουστικά στοιχεία $Y_p(k, n)$

Masking (Binary vs. Soft)

- Masking (art), from Wikipedia
  - In art, craft, and engineering, **masking** is the use of materials to protect areas from change, or to focus change on other areas. ….The term is derived from the word "mask", in the sense that it hides the face from view.

- HPSS:
  - Αναφέρεται στην αποκλειστική επιλογή του κυρίαρχου στοιχείου (harmonic or percussive) αγνοώντας την παρουσία του δεύτερου. Γίνεται με δύο τρόπους:
    - Binary or Hard Masking

$$\mathcal{M}_h(m,k) := \begin{cases} 1 & \text{if } \tilde{\mathcal{Y}}_h(m,k) \geq \tilde{\mathcal{Y}}_p(m,k) \\ 0 & \text{else} \end{cases}$$

$$\mathcal{M}_p(m,k) := \begin{cases} 1 & \text{if } \tilde{\mathcal{Y}}_p(m,k) > \tilde{\mathcal{Y}}_h(m,k) \\ 0 & \text{else.} \end{cases}$$

    - Soft Masking or Weiner Filtering
      - ε > 0 για να αποφευχθεί η διαίρεση με μηδέν

$$\mathcal{M}^{\mathrm{h}}(n,k) := \frac{\tilde{\mathcal{Y}}^{\mathrm{h}}(n,k) + \varepsilon/2}{\tilde{\mathcal{Y}}^{\mathrm{h}}(n,k) + \tilde{\mathcal{Y}}^{\mathrm{p}}(n,k) + \varepsilon},$$

$$\mathcal{M}^{\mathrm{p}}(n,k) := \frac{\tilde{\mathcal{Y}}^{\mathrm{p}}(n,k) + \varepsilon/2}{\tilde{\mathcal{Y}}^{\mathrm{h}}(n,k) + \tilde{\mathcal{Y}}^{\mathrm{p}}(n,k) + \varepsilon}$$

Signal Reconstruction

- Δημιουργία Σημάτων από το Αρμονικό και το Κρουστικό Φάσμα χρησιμοποιούνται:
  - Ο STFT **<u>του αρχικού σήματος</u>**
  - Τα δύο Masks
    - x_h = ISTFT $\{M_h(k, n)^* X(k, n)\}$
    - x_p = ISTFT $\{M_p(k, n)^* X(k, n)\}$
  - Με άλλα λόγια, τα $Y_h(k, n)$, $Y_p(k, n)$ χρησιμοποιούνται μόνο για τον υπολογισμό της μάσκας
- Προβληματισμοί:
  - Κατά πόσο ισχύει $x = x_h + x_p$?
    - Συχνά δεν ισχύει διότι υπάρχουν φαινόμενα παρεμβολής φάσης (phase interference) ανάμεσα στο κρουστικό και το αρμονικό σήμα
    - Ο STFT δεν είναι πάντοτε αναστρέψιμος
      - Κυρίως λόγω παραθυροποίησης και ολίσθησης

Εργαστηριακό Μέρος

1. Να υπολογισθούν α) ο διάμεσος (median) και ο β)μέσος όρος (mean) σε πίνακα μίας διάστασης και σε πίνακα δύο διαστάσεων.
2. Να διερευνηθεί το φίλτρο μέσου σε πίνακα δύο διαστάσεων, τόσο στην οριζόντια όσο και στην κάθετη διάσταση με χρήση της scipy.ndimage.median_filter
3. Να εφαρμοσθεί βήμα-βήμα ο αλγόριθμος HPSS στο αρχείο ‘input_audio/beat_jazz.wav’

Άσκηση

1. Χρησιμοποιώντας την υλοποίηση HPSS που είδαμε στο εργαστήριο, να εξετάσετε κατά πόσο το άθροισμα των σημάτων που προκύπτουν από HPSS ταυτίζεται με το αρχικό σήμα
2. Χρησιμοποιήστε τη βιβλιοθήκη librosa για να υπολογίσετε το αρμονικό και το κρουστικό στοιχείο του ίδιου σήματος (librosa.decompose.hpss)
3. Σχολιάστε

Παραπομπές

- Meinard Müller, Fundamentals of Music Processing, Spinger 2015
  - Κεφάλαιο 8 - Musically Informed Audio Decomposition
- M. Müller, Harmonic Percussive Source Separation, Course 2016
  - Lab Course 2016, Friedrich-Alexander-Universitat Erlangen-Nürenberg
- COLA – Constant Overlap Add Constraint
  - https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.check_COLA.html#scipy.signal.check_COLA