Musical Instrument Detection via Source Separation
Ψηφιακή Επεξεργασία Σήματος: Ερευνητική Εργασία

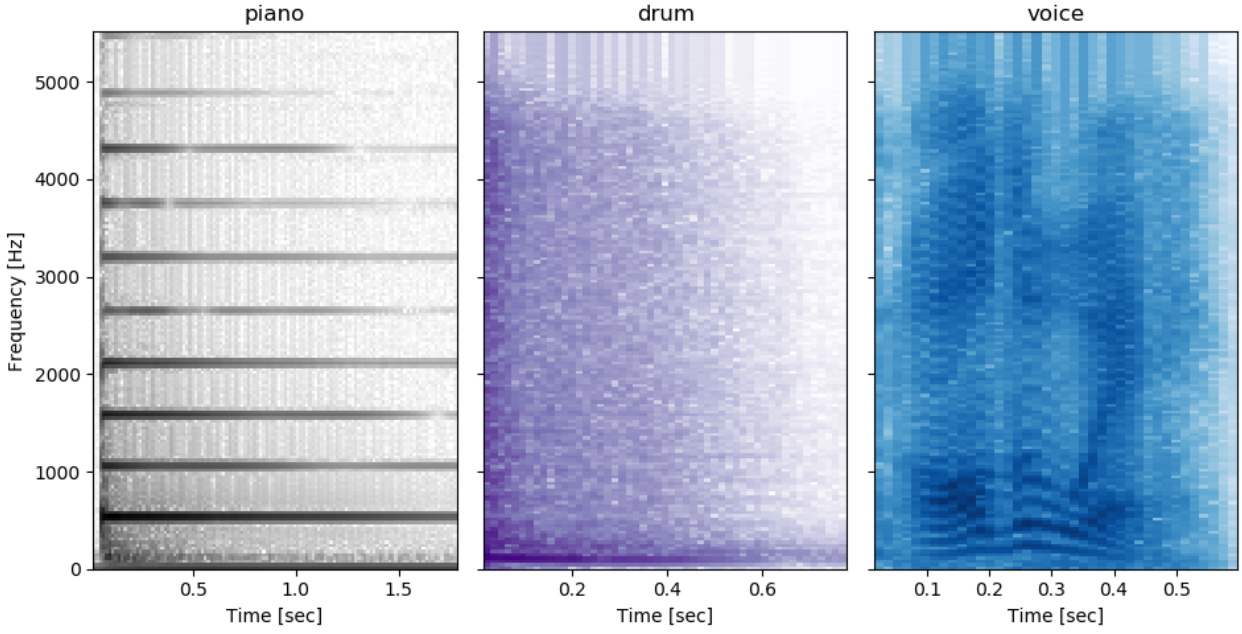
Δημήτριος Μπράλιος - 03116126 - dimitris.bralios@gmail.com

Εισαγωγή

Στην εργασία αυτή μελετάμε το πρόβλημα του Source Separation σε μουσικά σήματα με σκοπό το Musical Instrument Detection. Συγκεκριμένα, προσπαθούμε να διαχωρίσουμε μια παρατήρηση του μείγματος στα επιμέρους μουσικά σήματα του κάθε οργάνου. Αρχικά, εξετάζουμε τα χαρακτηριστικά των μουσικών οργάνων επικεντρώνοντας την προσοχή μας στις ειδοποιούς διαφορές που αποτελούν το ηχόχρωμα. Στην συνέχεια μελετάμε θεωρητικά αλγόριθμους που έχουν εφαρμοστεί πάνω στο δεδομένο πρόβλημα, εστιάζοντας στον αλγόριθμο NMF και τις παραλλαγές του. Πραγματοποιούμε μια σειρά από πειράματα σε απλά μουσικά μείγματα, συνδυασμοί νότων, και σχολιάζουμε τα αποτελέσματα με βάση την θεωρητική μελέτη που έχει προηγηθεί. Τέλος, εξετάζουμε τρόπους να εξάγουμε λάθος ανακατασκευής μεταξύ των ανακατασκευασμένων σημάτων και των αρχικών.

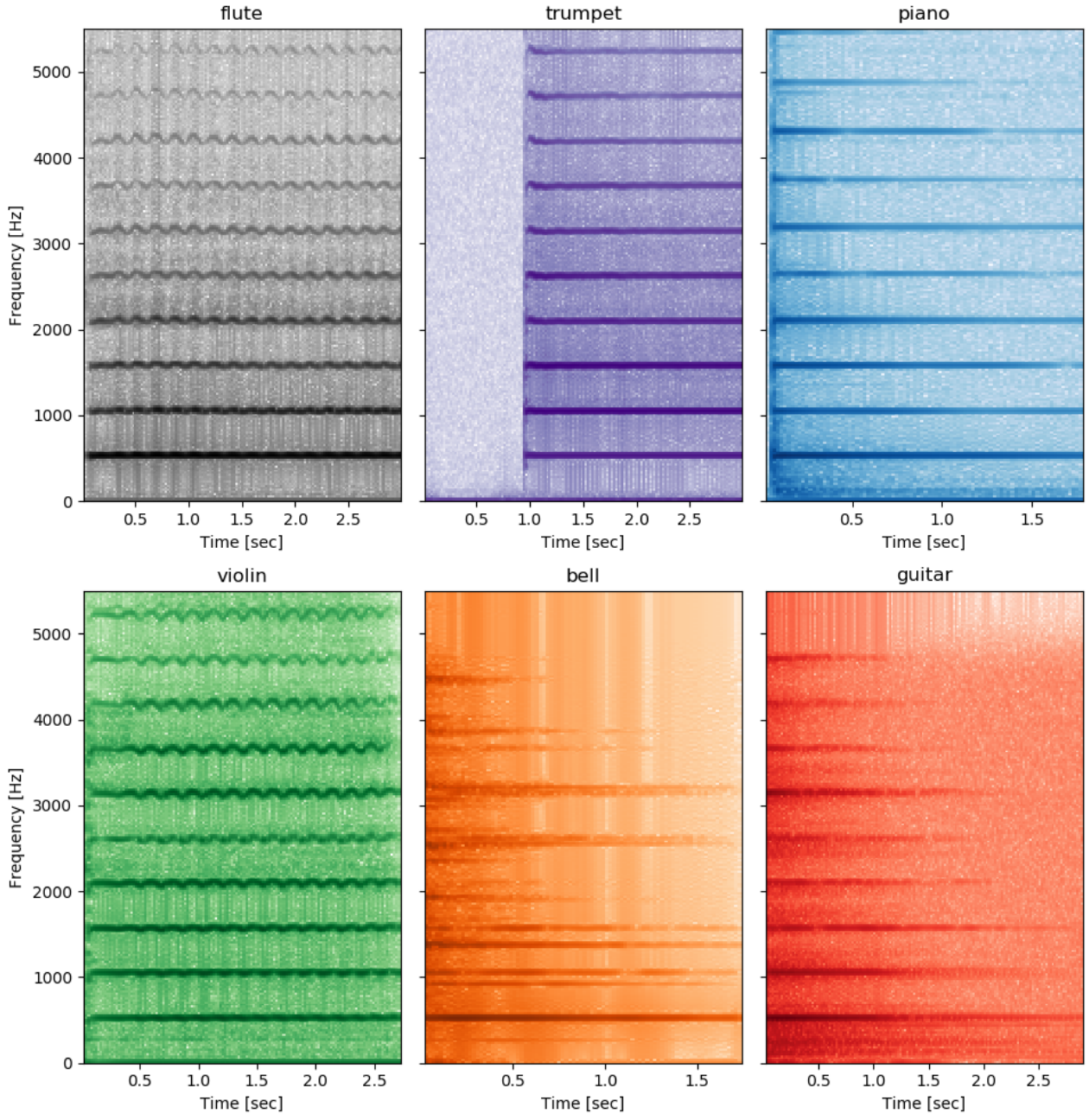
Χαρακτηριστικά Μουσικών Οργάνων

Ξεκινάμε μελετώντας τα χαρακτηριστικά των μουσικών οργάνων, εντοπίζοντας στοιχεία που τα διαφοροποιούν από άλλους ήχους αλλά και μεταξύ τους. Μια απλή κατηγοριοποίηση, των επιμέρους πηγών των μουσικών σημάτων, είναι σε κυρίως αρμονικές (harmonic), κυρίως κρουστικές (precussive) και σε φωνή [1]. Οι αρμονικές πηγές, παράγουν ήχους που περιέχουν ένα σύνολο τόνων (διακριτών συχνοτήτων), με πιο σημαντική την θεμελιώδη συχνότητα που καθορίζει το pitch. Αντίθετα, οι κρουστικές πηγές παράγουν σήματα, με ενέργεια σε μεγάλο εύρος συχνοτήτων, περιορισμένη χρονική διάρκεια και κρουστικά χαρακτηριστικά. Στην πραγματικότητα, τα περισσότερα όργανα παράγουν ήχους με αρμονικά και κρουστικά χαρακτηριστικά. Στο παρακάτω σχήμα βλέπουμε τα spectrograms, μιας νότας πιάνου, ενός χτύπηματος τυμπάνου και μια σύντομη φράση.



Η πλειοψηφία των οργάνων είναι σχεδιασμένη ώστε να δίνει στον καλλιτέχνη την δυνατότητα να παράγει αρμονικούς ήχους, με την επιθυμητή θεμελιώδη συχνότητα, την οποία τις περισσότερες φορές ταυτίζουμε με το pitch. Το ανθρώπινο αυτί αντιλαμβάνεται ήχους με λόγο θεμελιωδών συχνοτήτων 2:1 (οκτάβα) ως όμοιους, δηλαδή ορίζεται λογαριθμική κλίμακα στην συχνότητα. Έτσι, ένα pitch χαρακτηρίζεται από τον αριθμό οκτάβας και από το chroma. Σύνθετοι ήχοι δημιουργούνται από την ταυτόχρονη παρουσία από νότες με διαφορετικά pitch. Ενώ οι "ευχάριστοι" ήχοι συνήθως έχουν απλό λόγο pitch και κοινές αρμονικές. [2]

Συνεπώς, διαφορετικά μουσικά όργανα μπορούν να παράγουν ήχους με το ίδιο pitch. Αυτό που τα διαφοροποιεί, ανεξάρτητα από την ένταση, είναι η χροιά ή ηχόχρωμα (timbre). Ένας ακροατής μπορεί να αναγνωρίσει το όργανο που παράγει τον ήχο με βάση το ηχόχρωμά του. Όμως, το χαρακτηριστικό αυτό δεν έχει στενό ορισμό ενώ πολλές προσπάθειες έχουν γίνει για να βρεθούν παράμετροι που το προσδιορίζουν. Οι παράμετροι αυτοί χωρίζονται σε στατικά και δυναμικά χαρακτηριστικά του φάσματος του σήματος. Μερικά στατικά χαρακτηριστικά περιλαμβάνουν, το συχνοτικό κέντρο βάρους του φάσματος, την δομή των συχνοτικών κορυφών και την κατανομή της ενέργειας στις αρμονικές. Ενώ κάποια δυναμικά χαρακτηριστικά (στον χρόνο κατά τα στάδια attack, sustain, decay) είναι, η διαμόρφωση του πλάτους της ενέργειας των συχνοτήτων στον χρόνο όπως το (tremolo), η σχετική μεταβολή της ενέργειας των αρμονικών στον χρόνο και η διαμόρφωση συχνότητας των αρμονικών στον χρόνο (vibrato). [2, 3]



Στο παραπάνω σχήμα βλέπουμε spectrograms από διάφορα όργανα που παίζουν την νότα C5 (523 Hz). Παρατηρούμε έντονες διαφορές ανάμεσα στο πιάνο με τις αρμονικές που φθίνουν γρήγορα σε σχέση με το βιολί που παρουσιάζει κάποιου είδους διαμόρφωσης συχνότητας (vibrato) στις αρμονικές. Ακόμη, παρατηρούμε ότι η καμπάνα δεν έχει αυστηρά καθορισμένες αρμονικές όπως τα υπόλοιπα όργανα αλλά περιέχει αρκετές επιπλέον πλευρικές αρμονικές. Όμως, συγκρίνοντας το βιολί με το φλάουτο, οι διαφορές τους είναι σχεδόν ανεπαίσθητες, όπως η διαφορετική κατανομή της ενέργειας στις αρμονικές. Το γεγονός αυτό καθιστά την αναγνώριση και τον διαχωρισμό των μουσικών οργάνων, σε ένα μείγμα, αρκετά δύσκολο.

Μέθοδοι για Source Separation

Το πρόβλημα που μελετάμε δηλαδή, ο διαχωρισμός ενός μουσικού σήματος στα επιμέρους σήματα των μουσικών οργάνων που το αποτελούν, ανήκει στην κατηγορία Blind Source Separation (BSS). Στην ενότητα αυτήν, εξετάζουμε θεωρητικά μεθόδους που έχουν χρησιμοποιηθεί στο πρόβλημα του Source Separation, εστιάζοντας στην μέθοδο Non Negative Matrix Factorization (NMF).

Independent Component Analysis

Μια προσέγγιση του προβλήματος είναι να θεωρήσουμε ότι τα επιμέρους σήματα και οι πηγές που τα παράγουν, είναι ανεξάρτητες μεταξύ τους. Οι μέθοδοι οι οποίοι μας δίνουν ανεξάρτητα επιμέρους σήματα από κάποιες παρατηρήσεις του μείγματος, γενικεύονται με το όνομα Independent Component Analysis (ICA).

Πρακτικά, χρησιμοποιούμε επαναληπτικές μεθόδους για να ελαχιστοποιήσουμε ή να μεγιστοποιήσουμε ένα κριτήριο κόστους το οποίο σχετίζεται με την non-Gaussianity. Αφού γνωρίζουμε ότι η κατανομή πιθανότητας ενός αθροίσματος από ανεξάρτητες τυχαίες μεταβλητές τείνει σε κατανομή Gauss. Άρα, ταυτίζουμε την ανεξαρτησία με την non-Gaussianity. Συνεπώς, θέλουμε να μεγιστοποιήσουμε το non-Gaussianity των πηγών που εντοπίζουμε, ώστε να είναι ανεξάρτητες [4].

Οι παραπάνω μέθοδοι έχουν χρησιμοπιηθεί σε προβλήματα BSS, όμως έχουν περιορισμούς [5]. Αρχικά, με την συμβατική ICA, δεν μπορούμε να ανακτήσουμε παραπάνω πηγές από όσες παρατηρήσεις του μείγματος διαθέτουμε, που στην περίπτωσή μας είναι μια. Τεχνικές όπως η ISA αντιμετωπίζουν τον αυτόν τον περιορισμό. Ακόμη, η υπόθεση ότι οι πηγές είναι στατιστικά ανεξάρτητες δεν είναι εύστοχη, αφού συχνά σε ένα κομμάτι τα όργανα παίζουν την ίδια μουσική σύνθεση ή δεν είναι τελείως ανεξάρτητα. Τέλος, η ICA υποθέτει ένα γραμμικό στατικό μοντέλο μίξης, δηλαδή τα σήματα συνδυάζονται με τρόπο ανεξάρτητο του χρόνου.

Non Negative Matrix Factorization

Μια από τις πιο δημοφιλείς μεθόδους για την αντιμετώπιση του προβλήματος είναι η NMF. Με την μέθοδο αυτήν παραγοντοποιούμε έναν μη αρνητικό πίνακα V σε γινόμενο δυο μη αρνητικών πινάκων W και H, χαμηλότερης τάξης, δηλαδή

$$V \approx WH \qquad V \in \mathbb{R}_+^{F\times T},\ W \in \mathbb{R}_+^{F\times K},\ H \in \mathbb{R}_+^{K\times T}$$

Η ιδιότητα των μη αρνητικών στοιχείων του W δίνει την δυνατότητα ερμηνείας των K στηλών του ως διανυσμάτων βάσης, ενώ οι γραμμές του H μπορούν να ερμηνευθούν ως κέρδη ή ενεργοποίηση της αντίστοιχης βάσης. Ακόμη, εφόσον το H έχει μη αρνητικά στοιχεία έχουμε μόνο ενισχυτική και όχι καταστρεπτική συμβολή. Συνεπώς, ερμηνεύουμε την αναπαράσταση του V ως την σύνθεση θεμελιωδών κομματιών, πράγμα το οποίο δεν θα μπορούσαμε να κάνουμε αν είχαμε αρνητικές τιμές και καταστροφική συμβολή [6]. Το γεγονός αυτό καθιστά την NMF κατάλληλη μέθοδο για προβλήματα BSS.

Επειδή το σήμα μουσικής είναι μια πραγματική χρονοσειρά, πρέπει να μετασχηματιστεί σε μια μη αρνητική αναπαράσταση. Επιθυμούμε να χρησιμοποιήσουμε έναν μετασχηματισμό για τον οποίο ισχύει, ακόμα και προσεγγιστικά, η προσθετική ιδιότητα. Επομένως, συχνά χρησιμοποιούνται μετασχηματισμοί που δίνουν ένα μέτρο ενέργειας σε πεδίο χρόνου-συχνότητας, όπως το μέτρο του Short Time Fourier Transform (STFT). Υποθέτουμε ότι για το μέτρο $|X|$ του STFT του μείγματος θα ισχύει το παρακάτω, όπου $|X_i|$ το μέτρο του STFT του σήματος κάθε πηγής.

$$|X| = \sum_i |X_i|$$

Η υπόθεση αυτή είναι προσεγγιστικά καλή όταν τα επιμέρους σήματα είναι αραιά στο πεδίο χρόνου-συχνότητας [7]. Όμως, η προσεγγίση αυτή αγνοεί την φάση του STFT, χρησιμοποιώντας την μόνο για ανακατασκευή, με αποτέλεσμα να μην εκμεταλλεύεται όλη την πληροφορία που περιέχει η κυματομορφή.

Άλλοι μετασχηματισμοί χρόνου-συχνότητας βραχέως χρόνου που έχουν χρησιμοποιηθεί είναι Cosine Transform, Constant Q και Wavelet Transform [7]. Ενώ, έχουν αναπτυχθεί αλγόριθμοι για την εκμάθηση ενός μετασχηματισμού μαζί με την διάσπαση σε γινόμενο μη αρνητικών πινάκων [8].

Προσέγγιση με NMF

Χρησιμοποιώντας την NMF έχουμε μια supervised προσέγγιση για το ζητούμενο πρόβλημα [9, 7]. Υποθέτουμε ότι στο μείγμα έχουμε διάφορες κλάσεις ήχων. Για κάθε τέτοια κλάση μαθαίνουμε ένα λεξικό, δηλαδή τον αντίστοιχο πίνακα W, από καθαρές ηχογραφήσεις της. Το μέγεθος του λεξικού καθορίζει πόσο καλά μπορεί να περιγράψει ήχους της κλάσεις, όμως υπάρχει ο κίνδυνος να είναι τόσο μεγάλο ώστε να μπορεί να περιγράφει και ξένες κλάσεις. Συνδυάζοντας όλα τα λεξικά που διαθέτουμε, μπορούμε να εξηγήσουμε το μείγμα ως γραμμικό συνδυασμό των στοιχείων των λεξικών, βρίσκοντας τον πίνακα H. Συνεπώς, έχουμε την δυνατότητα να ανακατασκευάσουμε το σήμα της κάθε κλάσης χρησιμοποιώντας το αντίστοιχο λεξικό W με τις αντίστοιχες στήλες του πίνακα H, εκτελώντας ουσιαστικά source separation. Ακόμη, ο πίνακας H υποδεικνύει πότε μια κλάση είναι ενεργή, εκτελώντας detection. Τέλος, εύκολα μπορούμε να επεκτείνουμε την παραπάνω τεχνική σε semi-supervised, κρατώντας τα λεξικά που έχουμε ήδη μάθει σταθερά στον πίνακα W, και μαθαίνοντας τις υπόλοιπες στήλες οι οποίες θα αποτελέσουν λεξικό για τις άγνωστες κλάσεις που βρίσκονται στο μείγμα. Παρατηρούμε ότι, η παραπάνω προσέγγιση υποθέτει ότι τα λεξικά είναι διαφορετικά μεταξύ τους ώστε να μην περιγράφουν ξένες κλάσεις. Όμως ακόμα και αν έχουν όμοια στοιχεία μπορούμε να διαχωρίσουμε τις πηγές, εξετάζοντας τα χρονικά τους χαρακτηριστικά στον πίνακα H. [7]

Αλγόριθμοι υλοποίησης

Αρχικά, το πρόβλημα $V \approx WH,\ W, H \geq 0$ έχει άπειρες λύσεις, ενώ το πρόβλημα της ισότητας έχει αποδειχθεί πως είναι NP-Hard [10]. Επιθυμούμε να ελαχιστοποιήσουμε μια συνάρτηση κόστους (divergence), που δεν είναι απαραίτητα απόσταση.

$$\min_{W,H\geq 0} D(V\|WH)$$

Σε όλα τα NMF προβλήματα η $D(V\|WH)$ είναι non-convex ως προς τα H και W, αν όμως η $D(x\|y)$ είναι convex ως προς το όρισμα y τότε το πρόβλημα είναι convex ως προς το H με δεδομένο το W και ανάποδα. Μια συνάρτηση κόστους μπορεί να είναι η Ευκλείδεια απόσταση, η οποία είναι αρκετά απλή, αλλά ακατάλληλη για δεδομένα υψηλής διάστασης [11] με αποτέλεσμα να μην είναι καλή επιλογή για το συγκεκριμένο πρόβλημα. Συχνά χρησιμοποιείται η συνάρτηση κόστους Kullback Leibler, που όπως και η Itakura–Saito αποτελεί ειδική περίπτωση της β-divergence. Το πλεονέκτημά της είναι η ευαισθησία της σε χαμηλές ενέργειες, προσομοιώνοντας καλύτερα το ανθρώπινο ακουστικό σύστημα [12]. Η επιλογή συνάρτησης κόστους εξαρτάται από τα δεδομένα, το πρόβλημα και τους περιορισμούς του [11].

Έχοντας υπ' όψιν τα παραπάνω, καταστρώνουμε μια επαναληπτική μέθοδο επίλυσης, χρησιμοποιούμε δηλαδή τον αλγόριθμο Block Coordinate Descent. Συγκεκριμένα στο i-οστό βήμα, βρίσκουμε το βέλτιστο $H^{(i)}$ με δεδομένο το $W^{(i-1)}$, ενώ βρίσκουμε το $W^{(i)}$ με δεδομένο το $H^{(i)}$. Με τον τρόπο αυτό λύνουμε ένα πιο εύκολο πρόβλημα, εύρεση του βέλτιστου H με δεδομένο το W, ή ανάποδα, για το οποίο χρησιμοποιούμε τον αλγόριθμο βελτιστοποίησης Majorization - Minimization. Βρίσκουμε δηλαδή μια συνάρτηση άνω φράγμα για την μετρική απόστασης, την οποία μπορούμε να ελαχιστοποιήσουμε σε κλειστή μορφή. Επομένως, εξάγουμε κλειστούς τύπους που δίνουν τα H και W στο i-οστό βήμα. Στην συνέχεια, επαναλαμβάνουμε τα παραπάνω βήματα μέχρι να φτάσουμε σε μια ικανοποιητική λύση. Επισημαίνουμε ότι, η μέθοδος αυτή γενικά δεν διαθέτει θεωρητικές διασφαλίσεις αλλά πρακτικά δίνει εύκολα ικανοποιητικές λύσεις, οι οποίες εξαρτώνται από την αρχικοποίηση των πινάκων [13, 7].

Συγκεκριμένα για την συνάρτηση κόστους Kullback Leibler, αποδεικνύεται ότι βρίσκουμε τοπικά βέλτιστη λύση και καταλήγουμε στους τύπους, για την ανανέωση των W και H των Lee και Seung [13]. Η παραπάνω προσέγγιση μπορεί να επεκταθεί θέτοντας περιορισμούς ως προς την χρονική συνέχεια και την αραιότητα του πίνακα ενεργοποίησής H. Οι περιορισμοί προστίθενται ως όροι κανονικοποίησης στην συνάρτηση κόστους, χωρίς μεγάλες αλλαγές στον παραπάνω αλγόριθμο. Τροποποιούμε τον αλγόριθμο βρίσκοντας μια νέα Majorizing συνάρτηση που ελαχιστοποιείται σε κλειστή μορφή.

Περιορισμοί χρονικής συνέχειας και αραιότητας

Βασιζόμενος στα προηγούμενα ο Virtanen χρησιμοποιεί την παρακάτω συνάρτηση κόστους και κανόνες ανανέωσης των H και W [12], προσθέτοντας περιορισμούς για την χρονική συνέχεια και την αραιότητα.

$$c(W,H) = c_r(W,H) + \alpha c_t(H) + \beta c_s(H)$$

Η συνάρτηση κόστους αποτελέιται από το κόστος ανακατασκευής, το κόστος χρονικής συνέχειας και το κόστος αραιότητας. Η συνεισφορά των δυο τελευταίων καθορίζεται από τους παράγοντες α και β. Ως κόστος ανακατασκευής χρησιμοποιούμε την συνάρτηση Kullback Leibler.

$$c_r(W,H) = \sum_{f,\ t} \left([V]_{f,t} log \frac{[V]_{f,t}}{[WH]_{f,t}} - [V]_{f,t} + [WH]_{f,t} \right)$$

Έπειτα, έχουμε το κόστος χρονικής συνέχειας, που εξαρτάται από την ευκλείδεια απόσταση δύο διαδοχικών τιμών της αντίστοιχης γραμμής του πίνακα ενεργοποίησης H. Κανονικοποιούμε με την τυπική απόκλιση κάθε γραμμής $\sigma_k = \sqrt{\frac{1}{T}\sum_{t=1}^{T}[H]_{k,t}^2}$

$$c_t(H) = \sum_{k=1}^{K} \frac{1}{\sigma_k^2} \sum_{t=2}^{T} \left([H]_{k,t} - [H]_{k,t-1} \right)^2$$

Το κόστος αραιότητας είναι το εξής

$$c_s(H) = \sum_{k=1}^{K} \sum_{t=1}^{T} \left| \frac{[H]_{k,t}}{\sigma_k} \right|$$

Ο κανόνας ανανέωσης για το W είναι ο παρακάτω, ο οποίος αποδεικνύεται πως δεν αυξάνει το κόστος [13]. Όπου το $\odot$ συμβολίζει γινόμενο στοιχείο προς στοιχείο.

$$W \leftarrow W \odot \frac{\frac{V}{WH} H^T}{1H^T}$$

Για το H ισχύει ο παρακάτω πολλαπλασιαστικός κανόνας ανανέωσης, ο οποίος όμως δεν εγγυάται την μείωση του κόστους [12].

$$H \leftarrow H \odot \frac{\nabla c^-(W,H)}{\nabla c^+(W,H)}$$

Όπου $\nabla c(W,H) = \nabla c^+(W,H) - \nabla c^-(W,H)$ είναι η κλίση της συνάρτησης κόστους ως προς τον πίνακα H, την οποία χωρίζουμε σε θετικά και αρνητικά στοιχεία.

Έχοντας βρει τους πίνακες W και H μπορούμε να ανακατασκευάσουμε το σήμα ή κάποια συνιστώσα του. Πρώτα κατασκευάζουμε το spectrogram του μέτρου από τις στήλες του W και τις αντίστοιχες γραμμές του H που μας ενδιαφέρουν. Στην συνέχεια χρησιμοποιούμε την φάση του αρχικού spectrogram X. Έχοντας πλέον το μιγαδικό spectrogram μπορούμε να εφαρμόσουμε τον αντίστροφο STFT και να πάρουμε το σήμα στο πεδίο του χρόνου. Η τεχνική αυτή δεν έχει ισχυρή θεωρητική υποστήριξη αλλά παρέχει ικανοποιητικά αποτελέσματα. Εναλλακτικά, μπορούμε να εξάγουμε το spectrogram μέτρου της κάθε πηγής φιλτράροντας το αρχικό ως εξής:

$$|X_s| = \frac{W_s H_s}{\sum_{i=1}^{K} W_i H_i} \odot |X|$$

Πειραματικά αποτελέσματα δείχνουν ότι οι περιορισμοί αυτοί αυξάνουν, σε μικρό βαθμό, την ακρίβεια σωστής ανίχνευσης απλών ήχων νότων από διάφορα μουσικά όργανα, σε σχέση με την NMF με συνάρτηση κόστους Kullback Leibler [12].

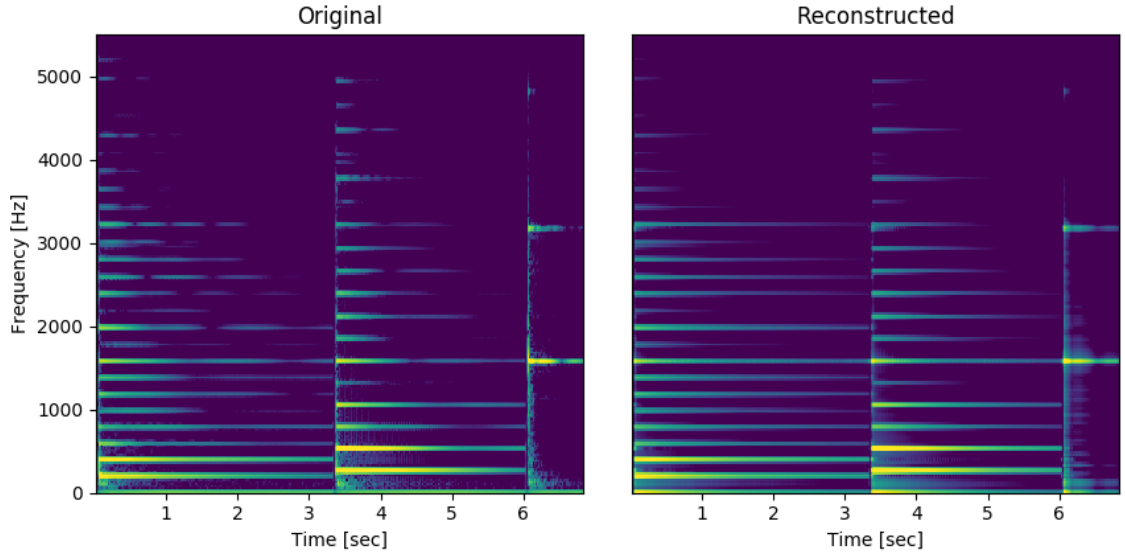
Πειραματικά Αποτελέσματα

Παρακάτω έχουμε μερικά παραδείγματα στα οποία εφαρμόζουμε την μέθοδο NMF σε απλά μουσικά σήματα, τα οποία αποτελούνται από νότες διαφόρων μουσικών οργάνων. Εξετάζουμε κάτω από ποίες συνθήκες μπορούμε να έχουμε διαχωρισμό οργάνων και ποία χαρακτηριστικά του ηχοχρώματος μπορούν να εκφραστούν

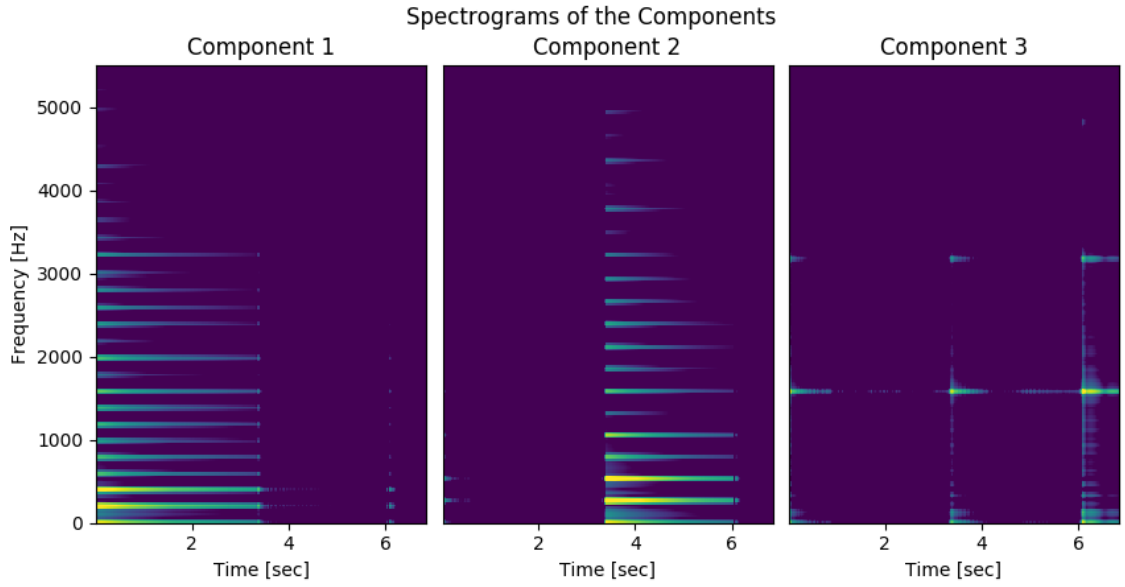
από μια περιγραφή διανύσματος βάσης και συνάρτηση χρονικής ενεργοποίησης. Έπειτα, μελετάμε μείγματα με χρονική επικάλυψη των μουσικών γεγονότων. Τέλος, συγκρίνουμε την απλή μέθοδο NMF έναντι της NMF με περιορισμούς, όπως εκφράστηκε στην προηγούμενη ενότητα. Εφαρμόζουμε τον απλό αλγόριθμο NMF, εκτός από το τελευταίο παράδειγμα που εφαρμόζουμε τον αλγόριθμο NMF με περιορισμούς. Κατασκευάζουμε τον πίνακα V χρησιμοποιώντας το μέτρο του STFT με παράθυρο hamming μήκους 40ms και επικάλυψη 50%.

Πείραμα 1: Νότες πιάνου χωρίς επικάλυψη

Στο πρώτο παράδειγμα, έχουμε τις νότες G3 (196 Hz), C4 (262 Hz) και G6 (1568 Hz) σε πιάνο χωρίς να επικαλύπτονται χρονικά. Στο πρώτο γράφημα βλέπουμε το spectrogram μέτρου του μείγματος στα αριστερά και στα δεξιά το ανακατασκευασμένο μείγμα, το οποίο μοιάζει αρκετά με το αρχικό με την διαφορά ότι είναι πιο ομαλό.

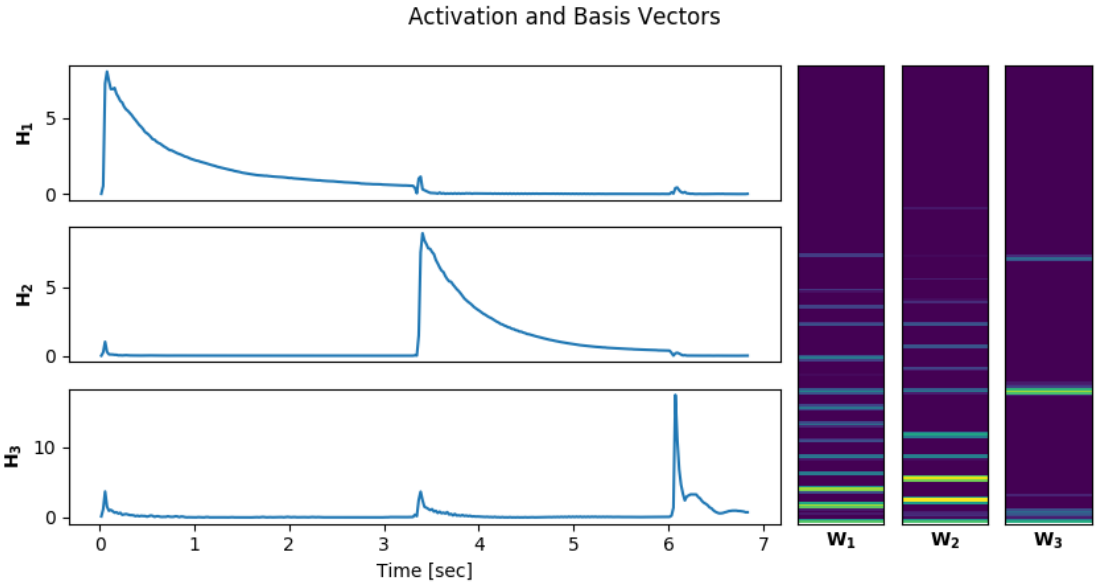


Στο επόμενο γράφημα βλέπουμε τα spectorgrams από τα επιμέρους σήματα που ξεχώρισε η μέθοδος. Το καθένα από αυτά αντιστοιχεί σε μια νότα, αλλά υπάρχουν και συμβολές και στις υπόλοιπες όπως βλέπουμε από το τρίτο σήμα που ενεργοποιείται σε κάθε νότα.



Στο τελευταίο γράφημα έχουμε τα διανύσματα βάσης στα δεξιά και τα αντίστοιχα διανύσματα ενεργοποίησης στα αριστερά. Παρατηρούμε καλύτερα από τις κορυφές των διανυσμάτων βάσεις ότι το καθένα ταιριάζει σε μια νότα. Το ίδιο συμπέρασμα βγάζουμε από τις κορυφές των διανυσμάτων ενεργοποίησης. Ακόμη βλέπουμε

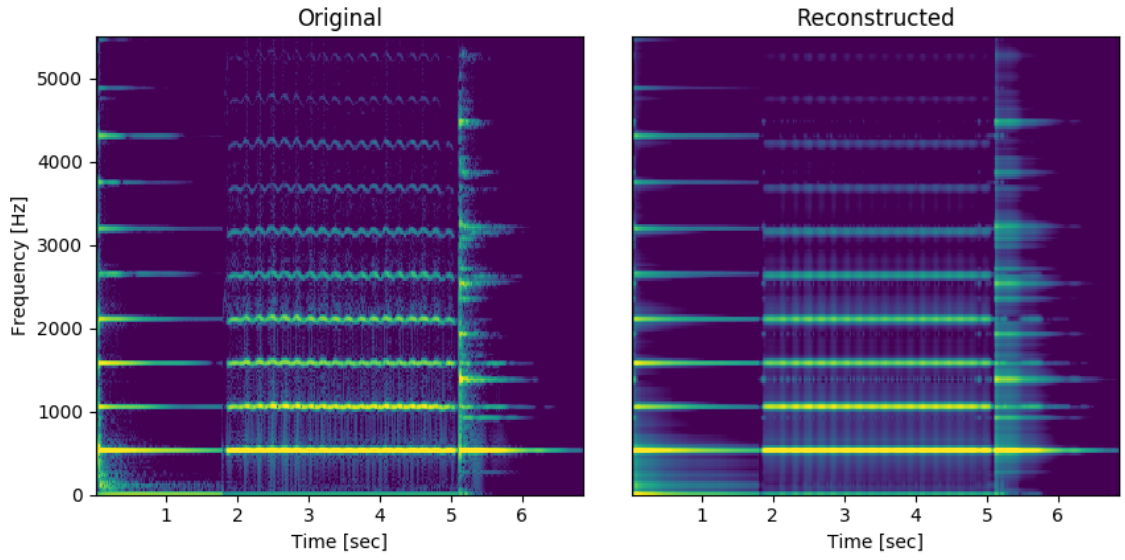
την μικρή συμβολή στις άλλες νότες κυρίως από την τρίτη βάση, πιθανώς επειδή έχει λίγες κορυφές και ταιριάζει εύκολα και στις υπόλοιπες.



Σε δυσκολότερες περιπτώσεις, όπως όταν έχουμε νότες με χρονική επικάλυψη τα αποτελέσματα δεν θα ήταν τα ίδια, δηλαδή τα διανύσματα βάσης δεν θα ήταν τόσο πιθανά να αναπαριστούν μια συγκεκριμένη νότα. Αντίθετα, τα διανύσματα βάσης θα αναπαριστούσαν γεγονότα όπως ο συνδυασμός δυο νότων. Συνεπώς, για να έχουμε αναπαράσταση με νότες θα πρέπει ο αλγόριθμος να έχει επαρκή δεδομένα ώστε να μπορεί να τις ξεχωρίσει. Τέλος, όταν ο αριθμός των βάσεων K που διαλέγουμε είναι διαφορετικός από τον αριθμό των νότων, τα αποτελέσματα εξαρτώνται από την επιλογή της συνάρτησης κόστους [14], ενώ προφανώς κάθε βάση δεν θα αναπαριστούσε μια νότα.

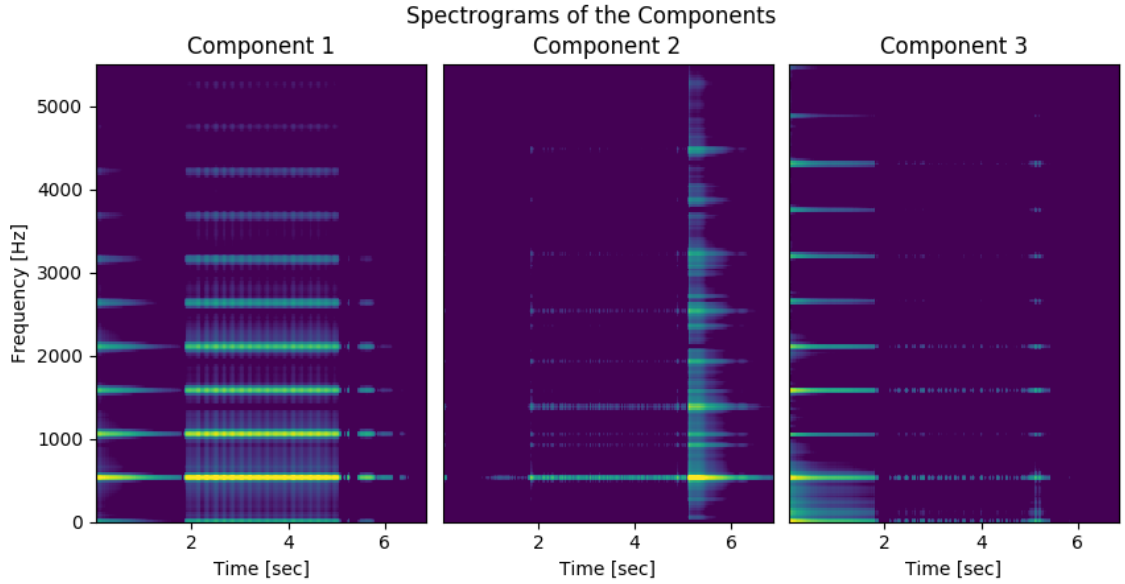
Πείραμα 2: Νότες από διαφορετικά μουσικά όργανα

Στο επόμενο παράδειγμα έχουμε την νότα C5 (523 Hz), από πιάνο, φλάουτο και καμπάνα. Στο spectogram του μείγματος, βλέπουμε τις διαφορές ανάμεσα στα μουσικά όργανα, όπως το vibrato στο φλάουτο και την διαρροή ενέργειας από τις αρμονικές στις γειτονικές συχνότητες στην καμπάνα. Το ανακατασκευασμένο σήμα, ειδικά το κομμάτι από το φλάουτο, φαίνεται ως μια χονδροειδής προσέγγιση χωρίς πολλή λεπτομέρεια.

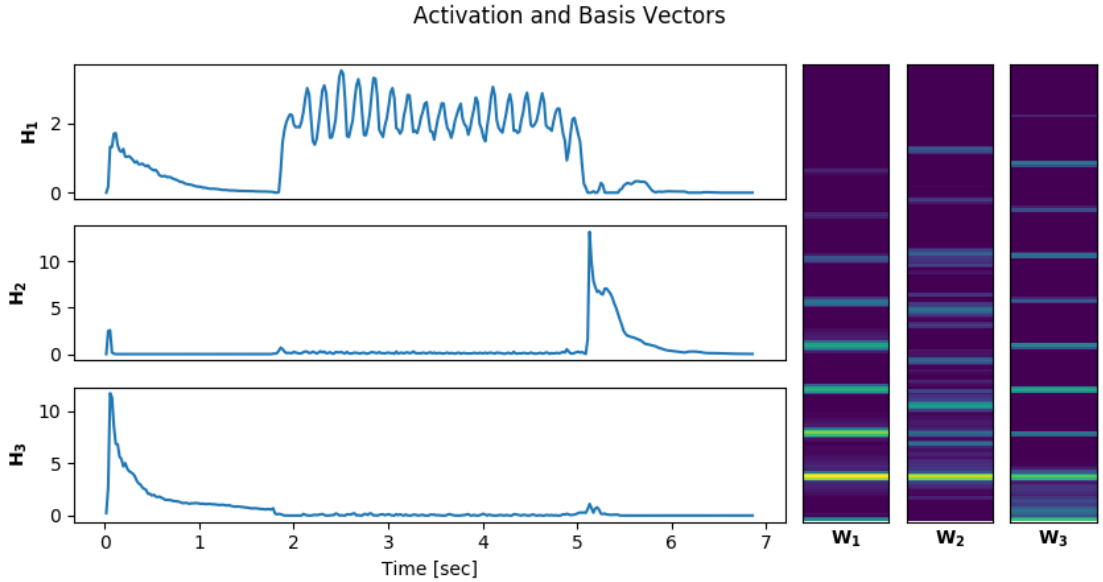


Βλέπουμε ότι τα επιμέρους σήματα αντιστοιχούν το καθένα σε ένα μουσικό όργανο, όμως οι συμβολές συμβαίνουν σε μεγαλύτερο βαθμό σε σχέση με πριν, όπως βλέπουμε από το πρώτο spectrogram. Επομένως,

δεν έχουμε καθαρή αναπαράσταση μουσικού οργάνου, αφού για την παραγωγή της κάθε νότας έχουμε είτε συνδυασμό βάσεων, είτε ενεργοποίηση βάσης που συμβάλει και σε άλλο όργανο.



Σε σχέση με το προηγούμενο παράδειγμα τα διανύσματα βάσης δεν έχουν τόσο απότομες κορυφές λόγω του διαφορετικού ηχοχρώματος του πιάνου με τα υπόλοιπα όργανα. Αξιοσημείωτο είναι ότι, η πρώτη βάση προσπαθεί να προσεγγίσει την την διαμόρφωση συχνότητας και πλάτους στις αρμονικές του φλάουτου, με μια διακύμανση της ενεργοποίησης της βάσης. Ακόμη, διακρίνουμε στο πρώτο διάνυσμα ενεργοποίησης την συμβολή στην νότα του πιάνου, ενώ από την τελευταία βάση παράγεται και το κομμάτι του attack, δηλαδή ο κρουστικός ήχος όταν το σφυρί χτυπάει την χορδή.

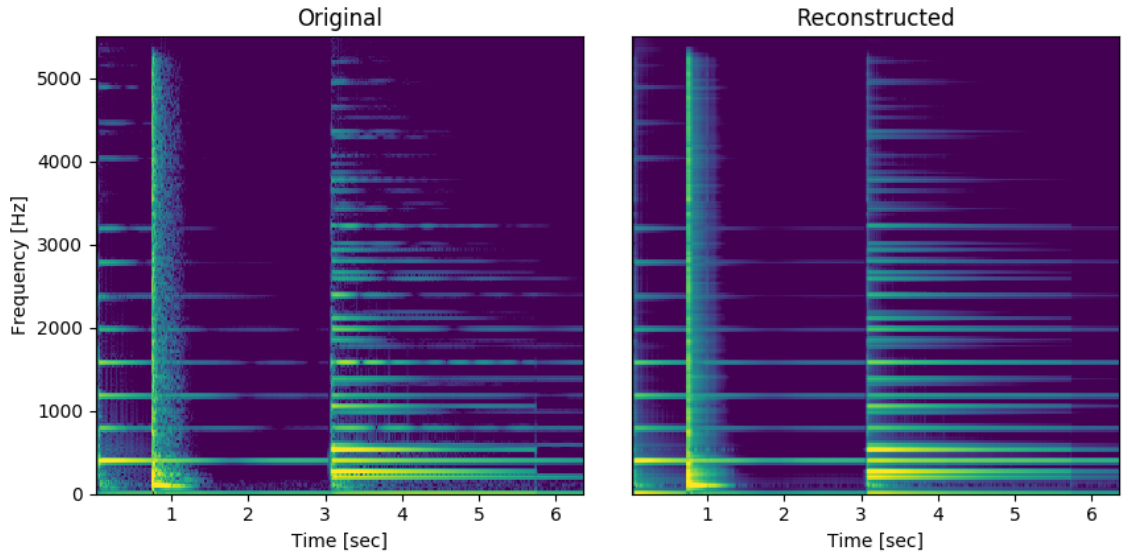


Ο αλγόριθμος μας δίνει αυτήν την περιγραφή επειδή τα διανύσματα βάσης μπορούν να αποδώσουν μόνο στατικά στοιχεία του φάσματος όπως κατανομή ενέργειας στις αρμονικές. Ενώ η μόνη ιδιότητα που είναι χρονικά μεταβαλλόμενη είναι η ενεργοποίηση της κάθε βάσης. Επομένως, δεν μπορούν να μοντελοποιηθούν επαρκώς τα χαρακτηριστικά του ηχοχρώματος, όπως η διαμόρφωση συχνότητας, με αποτέλεσμα να μην επαρκεί ακριβώς μια βάση για κάθε όργανο.

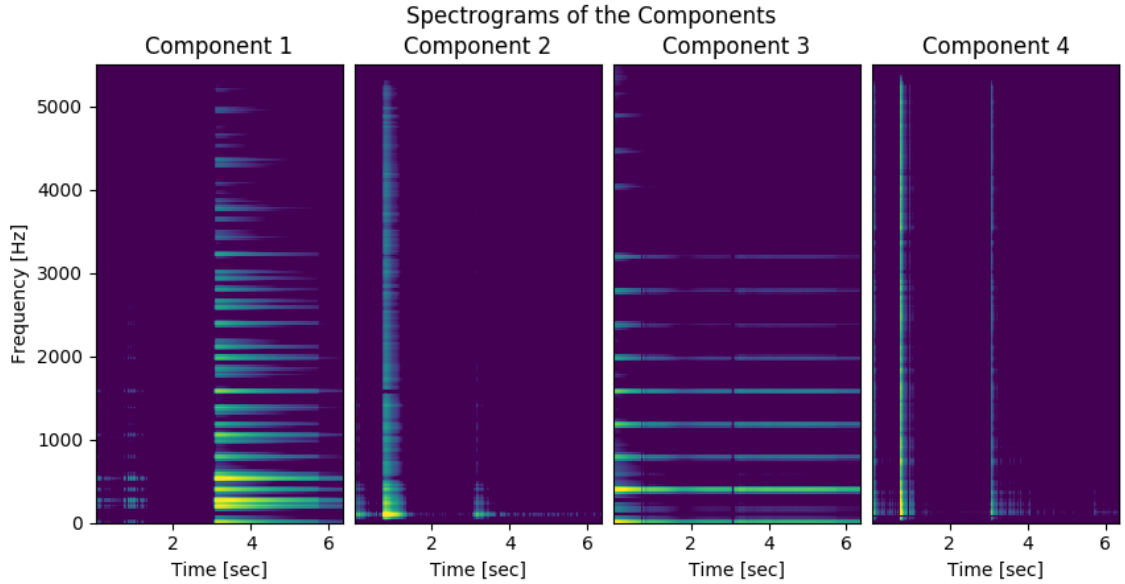
Όμως παρά τις δυσκολίες, έχουμε σε κάποιον βαθμό αντιστοίχηση βάσης με μουσικό όργανο, ενώ το ανακατασκευασμένο σήμα ακουστικά μοιάζει αρκετά στο αρχικό.

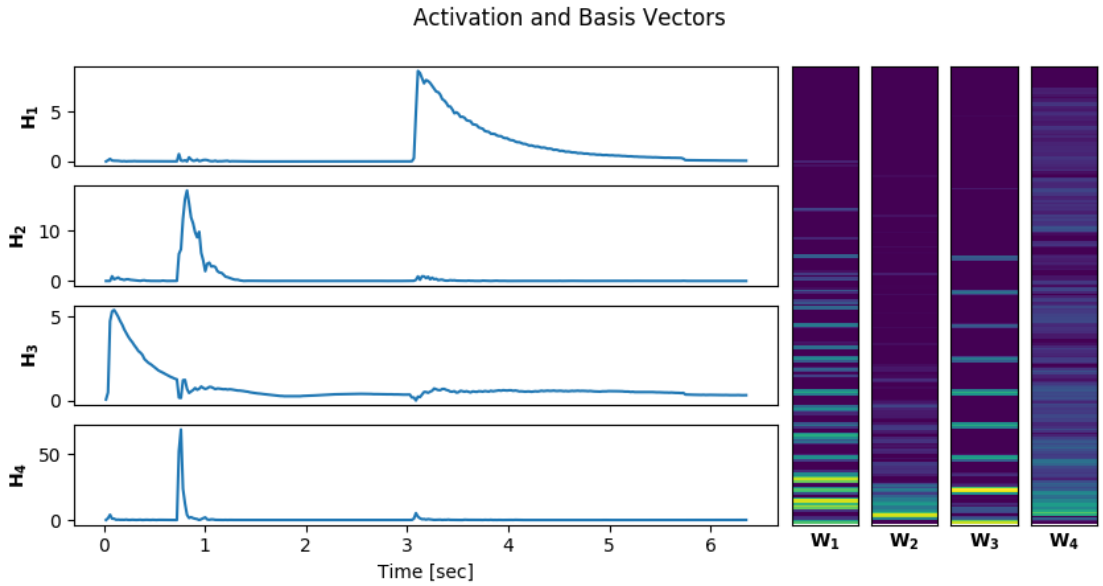
Πείραμα 3: Μουσικά γεγονότα με επικάλυψη

Στο επόμενο παράδειγμα έχουμε μείγμα με χρονική επικάλυψη μουσικών γεγονότων. Έχουμε, σε πιάνο, την νότα G4 (392 Hz) που επικαλύπτεται από ένα χτύπο τυμπάνου και στην συνέχεια τις νότες G3 (196 Hz) και C4 (262 Hz) που συμπίπτουν χρονικά.



Παρατηρώντας τα components βλέπουμε ότι έχουμε δυο που περιγράφουν το χτύπημα του τυμπάνου, ένα που περιγράφει μαζί τις νότες G3, C4 και ένα που περιγράφει την G4 καθώς και λιγότερο την G3.

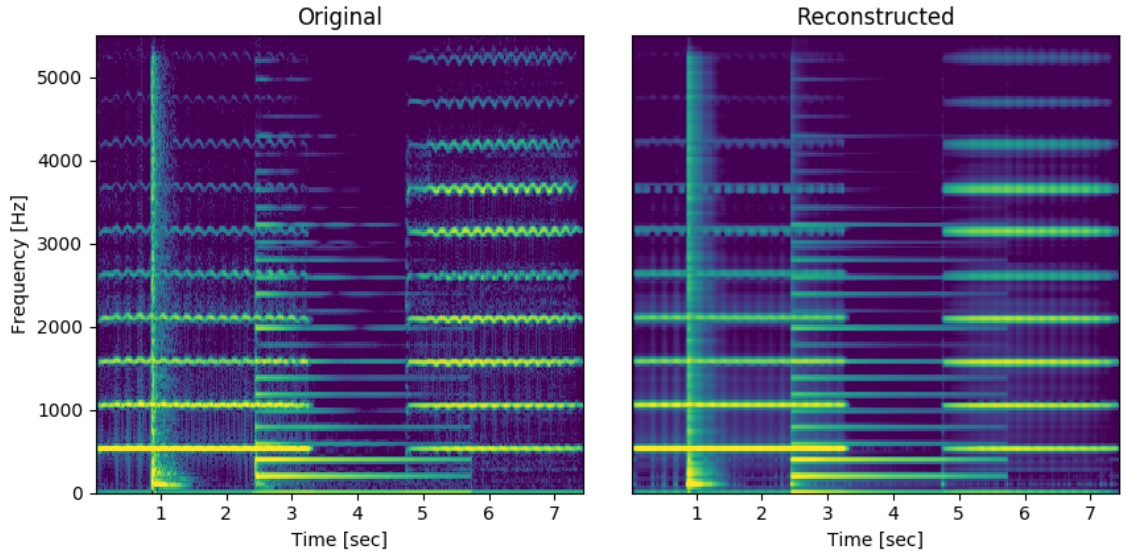


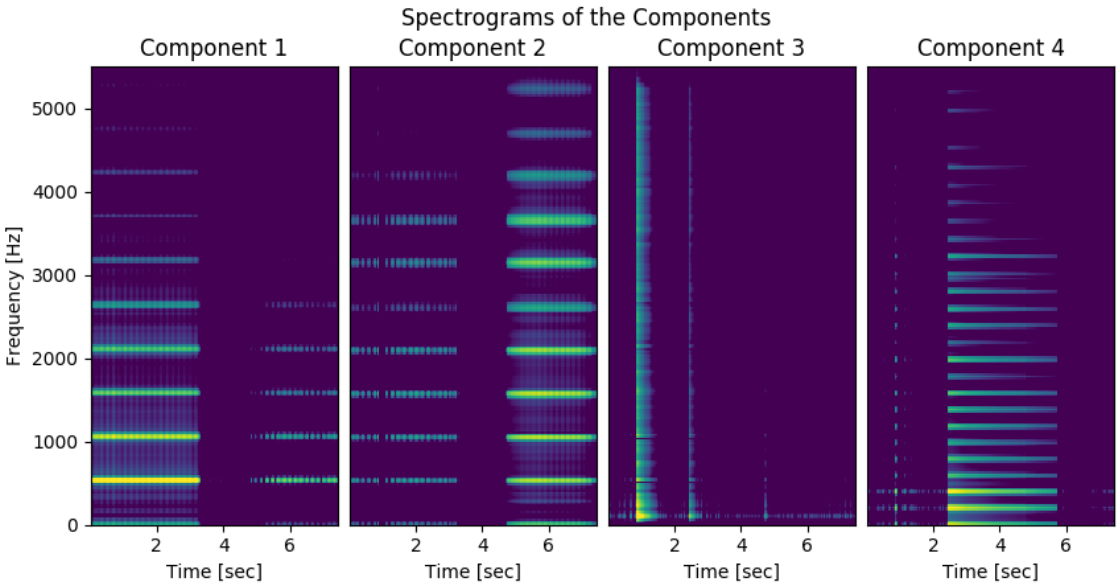


Επιβεβαιώνουμε τα παραπάνω παρατηρώντας τα διανύσματα ενεργοποίησης. Ο λόγος που η τρίτη βάση ενεργοποιείται και στις δυο τελευταίες νότες είναι το γεγονός ότι οι νότες G3 και G4 διαφέρουν κατά μια οκτάβα, με αποτέλεσμα να έχουν κοινές αρμονικές. Επίσης, όπως επισημάναμε η πρώτη βάση περιγράφει τον συνδυασμό των δυο ταυτόχρονων νότων, δηλαδή ένα μουσικό γεγονός. Αυτό συμβαίνει επειδή ο αλγόριθμος δεν διαθέτει επαρκή δεδομένα, με την κάθε νότα ξεχωριστά. Παρόλο που συναντάει την G4 που μοιράζεται αρμονικές με την G3, δεν μπορεί να ξεχωρίσει τις C4 και G3. Συνεπώς, ο αλγόριθμος εξάγει μοναδικά μουσικά γεγονότα και όχι κατ' ανάγκη νότες. Για να αναγνωρίσει μια νότα, θα πρέπει να παρέχουμε επαρκή δεδομένα είτε μεμονωμένες εμφανίσεις κάποιας νότας είτε διαφορετικούς συνδυασμούς που περιέχουν αυτήν [14].

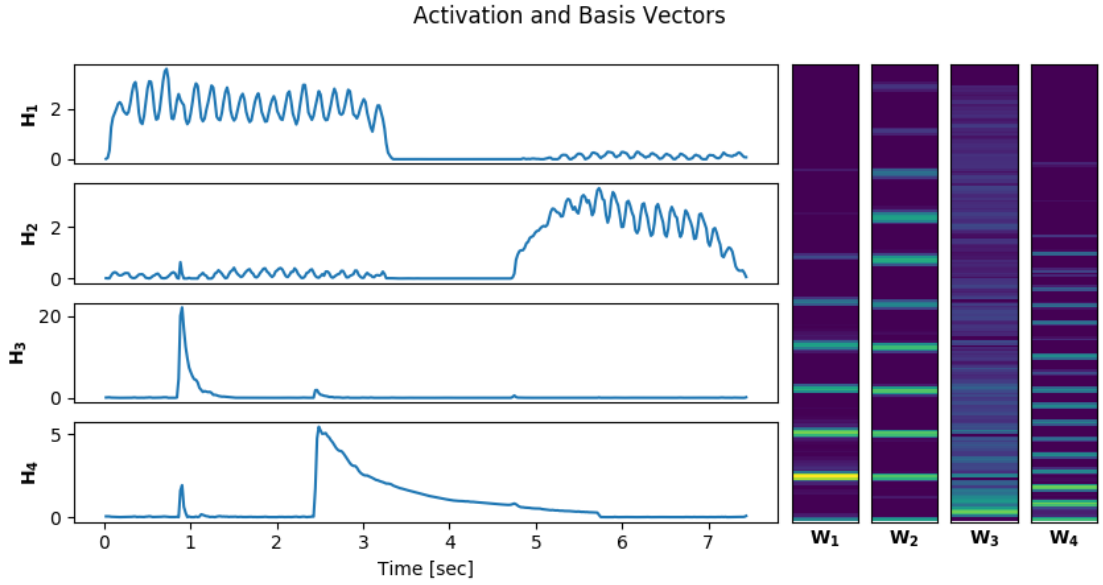
Πείραμα 4: Περιορισμοί χρονικής συνέχειας και αραιότητας

Στο τελευταίο παράδειγμα, εφαρμόζουμε τον αλγόριθμο NMF χρησιμοποιώντας την συνάρτηση κόστους με περιορισμούς χρονικής συνέχειας και αραιότητας. Συγκεκριμένα, χρησιμοποιούμε $\alpha = 100$ και $\beta = 10$. Το μείγμα αποτελείται από την νότα C5 (523 Hz) σε φλάουτο, ένα χτύπημα τυμπάνου, την νότα G3 (196 Hz) σε πιάνο και την νότα C5 (523 Hz) σε βιολί, όλα με μερική χρονική επικάλυψη.



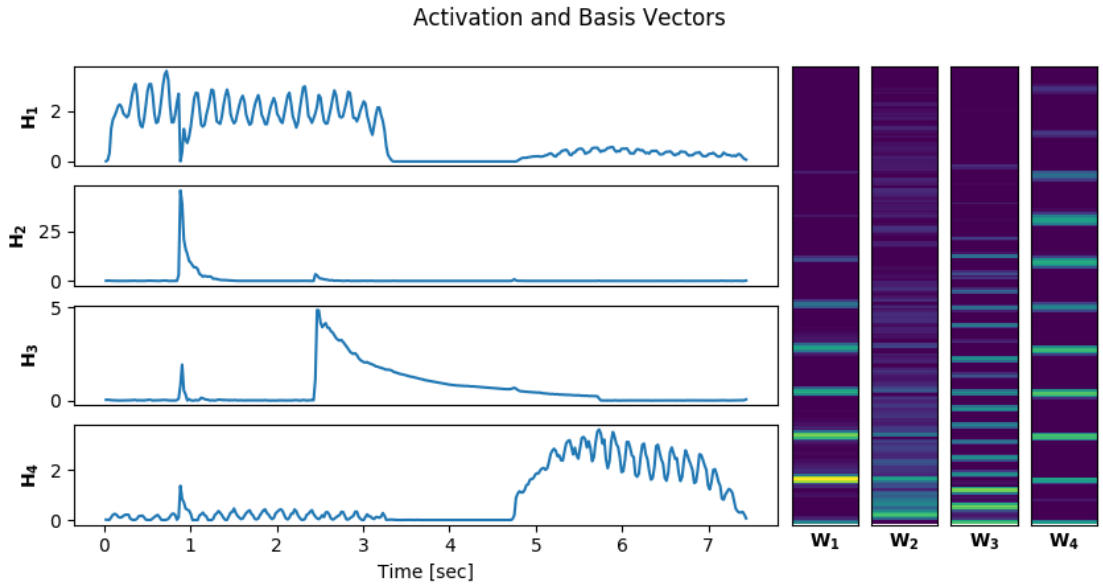


Το κάθε component που προκύπτει από την συγκεκριμένη λύση αντιστοιχεί κυρίως σε ένα όργανο, με κάποια συμμετοχή και στα υπόλοιπα. Παρατηρούμε πάλι τον τρόπο που προσεγγίζεται το vibrato και το tremolo στο φλάουτο και το βιολί με διακυμάνσεις της ενεργοποίησης. Ακόμη, έχουμε μια μικρή αμοιβαία συνεισφορά, αφού και τα δυο όργανα παίζουν την ίδια νότα.



NMF με περιορισμούς

Συγκρίνουμε την προηγούμενη λύση με μια λύση του αλγόριθμου NMF χωρίς επιπλέον περιορισμούς. Βλέπουμε ότι έτυχε τα διανύσματα βάσεων που προκύπτουν είναι πρακτικά τα ίδια με πριν. Διαφορές ανάμεσα στις δυο προσεγγίσεις εντοπίζουμε στα διανύσματα ενεργοποίησης. Λόγω του περιορισμού χρονικής συνέχειας τα διανύσματα ενεργοποίησης είναι πιο ομαλά και δεν έχουν τόσο απότομες μεταβολές. Σε συνδυασμό τον περιορισμό αραιότητας έχουμε ως αποτέλεσμα για παράδειγμα το χτύπημα του τυμπάνου να μην επηρεάζει τόσο τα υπόλοιπα components.



NMF χωρίς περιορισμούς

Τέλος, για να δούμε αν τα ανακατασκευασμένα components προσεγγίζουν καλύτερα τα αρχικά με την μέθοδο με περιορισμούς, εξετάζουμε μέτρα σύγκρισης και παραθέτουμε τα αποτελέσματα στην επόμενη ενότητα.

Αξιολόγηση Διαχωρισμού

Στην ενότητα αυτήν εξετάζουμε τρόπους για να αξιολογήσουμε τα ανακατασκευασμένα σήματα σε σχέση με τα αρχικά. Υποθέτουμε ότι τα αρχικά σήματα είναι ευθυγραμμισμένα χρονικά με την θέση τους στο μείγμα. Η σύγκριση πραγματοποιείται μεταξύ των spectrogram μέτρου, ώστε η ανακατασκευή με την χρήση της αρχικής φάσης να μην παίξει κάποιον ρόλο. Με Y συμβολίζουμε το αρχικό σήμα, και με $\hat{Y}$ το ανακατασκευασμένο. Τα μέτρα σύγκρισης που εξετάζουμε είναι τα εξής:

$$D_1 = \left\| \frac{\hat{Y}}{\|\hat{Y}\|} - \frac{Y}{\|Y\|} \right\|^2 \qquad D_2 = \frac{\|Y\|^2}{\|\hat{Y} - Y\|^2} \qquad D_3 = \left\| Y \odot log\left(\frac{Y}{\hat{Y}}\right) - Y + \hat{Y} \right\|^2$$

Το D_1 είναι το πιο απλό, αρχικά κανονικοποιούμε τα σήματα με την L_2 νόρμα τους και παίρνουμε το τετράγωνο της L_2 νόρμας της διαφοράς. Παίρνει τιμές από 0 έως 2 με αποτέλεσμα να μην έχει καλή ανάλυση όσο πλησιάζουμε στο 2 [15]. Το D_2 είναι ο σηματοθορυβικός λόγος και αποδίδει καλύτερά από το προηγούμενο αφού παίρνει τιμές από 0 έως $+\infty$ μεγαλώνοντας όσο καλύτερο είναι το ταίριασμα. Τέλος, το D_3 μοιάζει με την Kullback Leibler divergence. Επιπλέον με τα D_2, D_3 έχουμε την δυνατότητα να χρησιμοποιήσουμε κανονικοποιημένες τιμές των Y, $\hat{Y}$.

Η διαδικασία αξιολόγησης ακολουθεί τα εξής βήματα. Αρχικά, εκτελούμε τον αλγόριθμο NMF λαμβάνοντας τα επιμέρους σήματα. Έπειτα υπολογίζουμε για κάθε σήμα τα μέτρα σύγκρισης σε σχέση με κάθε αρχικό επιμέρους σήμα. Αντιστοιχίζουμε έτσι ένα ανακατασκευασμένο σήμα στο αρχικό component του μείγματος με το οποίο ταιριάζει καλύτερα. Εφαρμόζοντας το παραπάνω για τα παραδείγματα της προηγούμενης ενότητας όλες οι μετρικές συμπίπτουν στις αντιστοιχίσεις.

Επίδραση NMF με περιορισμούς στον διαχωρισμό

Για να συγκρίνουμε τις δυο παραλλαγές του αλγορίθμου, αντιπαραβάλλουμε τα μέτρα σύγκρισης των αρχικών σημάτων με τα αντίστοιχα ανακατασκευασμένα για τον κάθε αλγόριθμο. Συγκεκριμένα για το τέταρτο παράδειγμα της προηγούμενης ενότητας έχουμε τα παρακάτω αποτελέσματα.

Components	D_1	D_2	D_3
Flute	8.800e-02	1.161e+01	2.058e+08
Violin	2.791e-01	3.701e+00	1.986e+07
Drum	3.843e-01	2.234e+00	1.836e+07
Piano	**9.138e-02**	**1.110e+01**	1.298e+07

NMF χωρίς περιορισμούς

Components	D_1	D_2	D_3
Flute	**4.390e-02**	**2.286e+01**	**3.442e+07**
Violin	**1.704e-01**	**6.036e+00**	**1.053e+07**
Drum	**2.521e-01**	**4.091e+00**	**1.509e+07**
Piano	9.168e-02	1.107e+01	**1.194e+07**

NMF με περιορισμούς

Συγκρίνοντας τους δυο πίνακες συμπεραίνουμε ότι για τα σήματα από φλάουτο, βιολί και τύμπανο ο αλγόριθμος NMF με περιορισμούς αποδίδει καλύτερα σύμφωνα με όλες τις μετρικές. Αντίθετα, για το σήμα από το πιάνο αποδίδει καλύτερα ο αλγόριθμος χωρίς περιορισμούς σύμφωνα με τα D_1 και D_2. Χρησιμοποιώντας κανονικοποιημένες τιμές των Y, $\hat{Y}$ στα D_2 και D_3, δεν έχουμε διαφορές για το παράδειγμα που εξετάζουμε.

Επισημαίνουμε ότι για να βγάλουμε συμπέρασμα για την σχετική απόδοση των αλγορίθμων, πρέπει να επαναλάβουμε την άνωθεν διαδικασία σε μεγάλο όγκο δεδομένων. Όμως, το συγκεκριμένο παράδειγμα υποστηρίζει τον ισχυρισμό του Virtanen [12], ότι ο συγκεκριμένος αλγόριθμος βελτιώνει τον σηματοθορυβικό λόγο των ανακατασκευασμένων σημάτων.

Συμπεράσματα

Οι ιδιότητες που διαφοροποιούν τα μουσικά όργανα συγκαταλέγονται στο ηχόχρωμα, ενώ διαθέτουν και στατικά και δυναμικά χαρακτηριστικά. Η μέθοδος NMF δίνει εύκολα μια παραγοντοποίηση σε συνιστώσες που συμβάλουν μόνο προσθετικά, διαχωρίζοντας έτσι ένα μείγμα στα επιμέρους συστατικά του. Εφαρμόζοντας την σε μουσικά σήμα συμπεραίνουμε ότι, μια στατική βάση που προκύπτει, δεν μπορεί να περιγράψει όλους τους ήχους που παράγει ένα μουσικό όργανο, αλλά έχοντας επαρκή δεδομένα μπορεί να περιγράψει μια νότα ή κάποιον κρουστικό ήχο. Συνεπώς, για να πραγματοποιήσουμε ανίχνευση μουσικού οργάνου με τον συγκεκριμένο αλγόριθμο πρέπει να μελετήσουμε την βάση σε συνδυασμό με τα χρονικά χαρακτηριστικά της ενεργοποίησης. Η μέθοδος επεκτείνεται με την προσθήκη περιορισμών αραιότητας και χρονικής συνέχειας, χάνοντας κάποιες θεωρητικές διασφαλίσεις αλλά πρακτικά έχουμε λίγο καλύτερη απόδοση στον διαχωρισμό. Τέλος, ο αλγόριθμος NMF διαθέτει αρκετές επεκτάσεις [11] οι οποίες θα μπορούσαν να αποδώσουν καλύτερα τα χαρακτηριστικά του ηχοχρώματος αλλά πλέον τα μαθηματικά παύουν να είναι απλά και δεν γενικεύουν το ίδιο εύκολα [7].

Αναφορές

[1] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stöter, "Musical source separation: An introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, 2018.

[2] M. Muller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[3] M. J. Tramo, P. Cariani, A. Oxenham *et al.*, "Hst. 725 music perception and cognition, spring 2009," *MIT OpenCourseWare*, 2009.

[4] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[5] G. Clifford, "Blind source separation: principal & independent component analysis," *MIT OpenCourseWare*, 2007.

[6] P. Smaragdis, "About this non-negative business," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–1.

[7] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[8] D. Fagot, H. Wendt, and C. Févotte, "Nonnegative matrix factorization with transform learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2431–2435.

[9] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 414–421.

[10] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, no. 257, pp. 257–291, 2014.

[11] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[12] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[14] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*. IEEE, 2003, pp. 177–180.

[15] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.