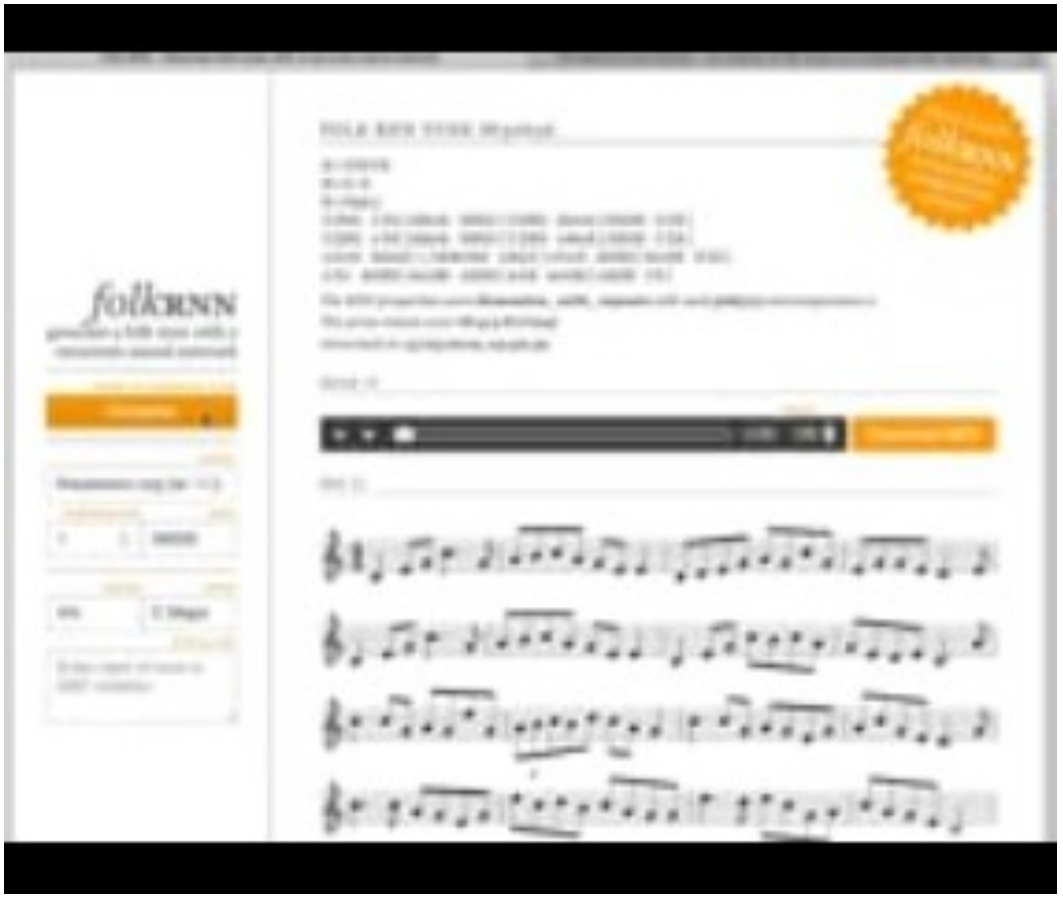
# O que significam esses 5.599.881 parâmetros?

Estevan Barbará

# RNN-Folk

- Rede Neural para geração de músicas “folk”, utilizando Redes de longa memória de curta duração (LSTM)
  - Proposta por (STURM et al, 2016)
  - Disponível em https://github.com/IraKorshunova/folk-rnn
  - Implementação acessível em https://folkrnn.org/



X:1

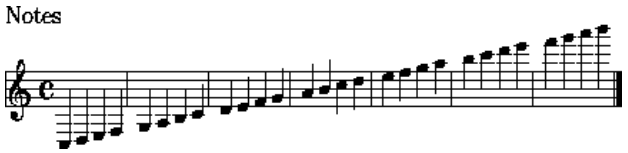
T:Notes

M:C

L:1/4

K:C

C, D, E, F,|G, A, B, C|D E F G|A B c d|e f g a|b c' d' e'|f' g' a' b'|]




# RNN-Folk – entradas e saídas

- Músicas em notação ABC
  - Descrita em http://abcnotation.com/
  - Padrão para diversos usos de música, particularmente o estilo “folk”
  - Composta por duas partes:
    - Cabeçalho de cinco linhas, contendo Id (X), Título (T), Compositor (C), Métrica (M), Duração padrão (L) e Tom (K). Cabe observar que apenas X, T e K são realmente obrigatórios, sendo o M padrão o compasso 4/4 e o L padrão a colcheia (1/8 de tempo).
    - Notas, em representação por letra
      - Letras maiúsculas (C B A) representam a oitava de baixo (logo abaixo da marcação numa clave de Sol)
      - Letras minúsculas (g f c) representam a oitava de cima
      - Uma vírgula após a nota indica que ela está uma oitava abaixo
      - Um apóstrofo indica uma oitava acima

X:1

T:Note lengths and default note length

M:C

K:C

L:1/16

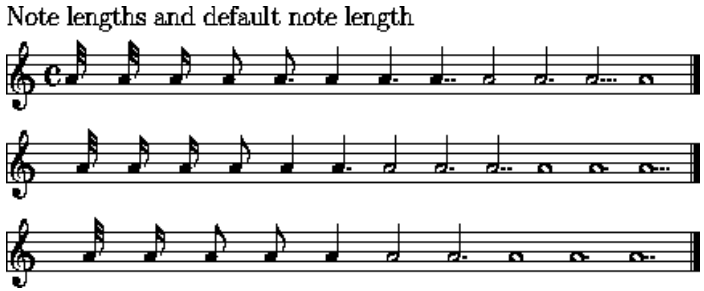
A/2 A/ A A2 A3 A4 A6 A7 A8 A12 A15 A16|]

L:1/8

A/4 A/2 A/ A A2 A3 A4 A6 A7 A8 A12 A15|]

L:1/4

A/8 A/4 A/2 A/ A A2 A3 A4 A6 A7|]

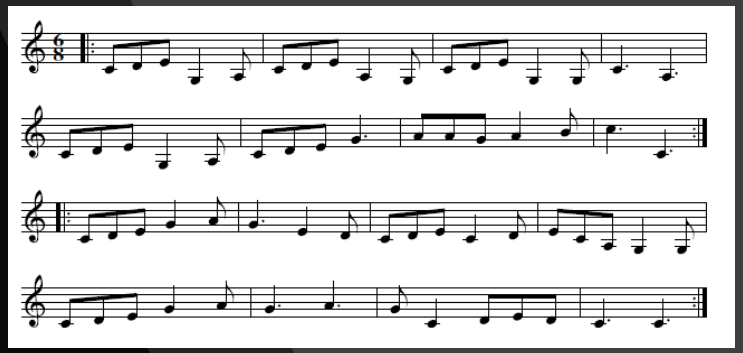


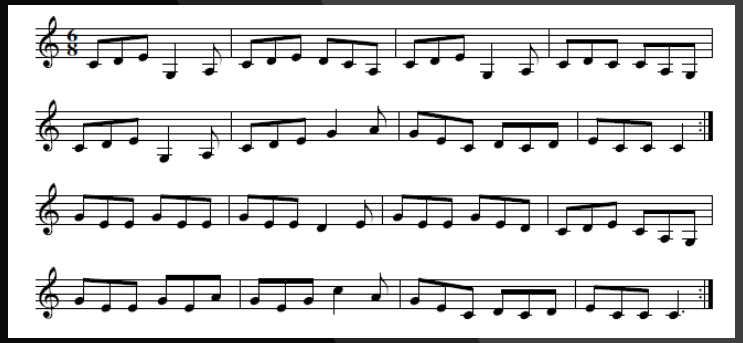

# RNN-Folk – entradas e saídas

- Os tempos são dados por uma numeração multiplicativa após a nota:
  - No exemplo, a primeira linha tem como padrão a semicolcheia (1/16). As notas da linha possuem os tempos:
    - Fusa (1/32 ou semicolcheia/2)
    - Fusa (a barra de divisão sozinha pode ser tratada como /2)
    - Semicolcheia (1/16)
    - Colcheia (1/8 ou semicolcheia x 2)
    - Colcheia + meio-tempo (3/16 ou semicolcheia * 3)
    - Semínima (1/4 ou semicolcheia * 4)
    - Semínima + Colcheia (3/8 ou semicolcheia * 6)
    - Semínima + Colcheia + Semicolcheia (7/16 ou semicolcheia * 7)
    - Mínima (1/2 ou semicolcheia * 8)
    - Mínima + Semínima (3/4 ou semicolcheia * 12)
    - Mínima + Semínima + Colcheia + Semicolcheia (15/16 ou semicolcheia * 15)
    - Semibreve (1 ou semicolcheia * 16)

# RNN-Folk – composição gerada

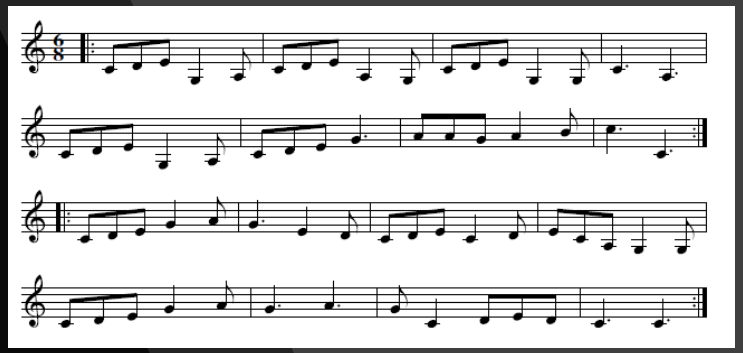
- Partindo de uma rede treinada com o dataset obtido em http://thesession.org
  - A figura de cima representa uma composição gerada pelo RNN-Folk
  - A figura de baixo representa a composição Páidín Ó Raifeartaigh (também repetida na música The Quaker's Wife).
    - Considereda a composição de treinamento mais similar à composição gerada

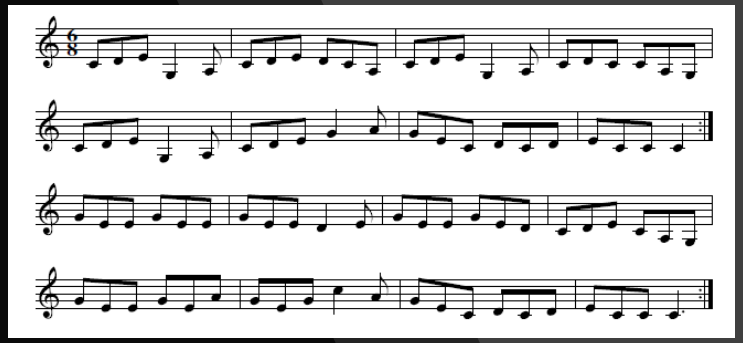




# RNN-Folk – composição gerada

- O resultado apresenta uma forma folk padrão
  - Forma AABB
  - 8 tempos em cada compass
  - Ambas as partes começam e terminam em notas tônicas, apresentam forte repetição e variação de uma melodia base
- No entanto, a composição não possui equivalente no dataset.
  - *De onde vieram as diferenças? Onde na rede estão codificadas as variações de tom, métrica, altura e duração das notas?*





# RNN-Folk – Arquitetura da rede

- A rede RNN-folk é composta por um layer de entrada, três layers LSTM intermediários e um SOFTMAX de saída
  - Os layers de entrada e de saída correspondem a um vetor binário de 137 dimensões, representando:
    - transcription (2);
    - meter (7);
    - mode (4);
    - measure (5);
    - pitch (85);
    - grouping (9);
    - and duration (25).

# RNN-Folk – Arquitetura da rede

- Layers internos:
  - Cada layer interno é composto por 512 Neurônios LSTM, cada um carregando a seguinte notação:
    - $I_t$: saída dos *input gates*
    - $F_t$: saída dos *forget gates*
    - $O_t$: saída dos *output gates*
    - $C_t$: saída dos *cell gates*
    - $H_t$: estado oculto do layer
    - $B_i$, $B_f$, $B_o$, $B_c$: Viéses associados aos *gates*
    - $W_{xi}$, $W_{xf}$, $W_{xo}$, $W_{xc}$: peso associado à entrada e aos *gates*
    - $W_{hi}$, $W_{hf}$, $W_{ho}$, $W_{hc}$: peso associado ao estado oculto e aos *gates*

$$\mathbf{i}_t^{(i)} \leftarrow g(\mathbf{W}_{xi}^{(i)}\mathbf{y}_t^{(i)} + \mathbf{W}_{hi}^{(i)}\mathbf{h}_{t-1}^{(i)} + \mathbf{b}_i^{(i)}) \quad (1)$$

$$\mathbf{f}_t^{(i)} \leftarrow g(\mathbf{W}_{xf}^{(i)}\mathbf{y}_t^{(i)} + \mathbf{W}_{hf}^{(i)}\mathbf{h}_{t-1}^{(i)} + \mathbf{b}_f^{(i)}) \quad (2)$$

$$\mathbf{o}_t^{(i)} \leftarrow g(\mathbf{W}_{xo}^{(i)}\mathbf{y}_t^{(i)} + \mathbf{W}_{ho}^{(i)}\mathbf{h}_{t-1}^{(i)} + \mathbf{b}_o^{(i)}) \quad (3)$$

$$\mathbf{c}_t^{(i)} \leftarrow \tanh(\mathbf{W}_{xc}^{(i)}\mathbf{y}_t^{(i)} + \mathbf{W}_{hc}^{(i)}\mathbf{h}_{t-1}^{(i)} + \mathbf{b}_c^{(i)}) \odot \mathbf{i}_t^{(i)} + \mathbf{f}_t^{(i)} \odot \mathbf{c}_{t-1}^{(i)} \quad (4)$$

$$\mathbf{h}_t^{(i)} \leftarrow \tanh(\mathbf{c}_t^{(i)}) \odot \mathbf{o}_t^{(i)}. \quad (5)$$

# RNN-Folk – arquitetura interna

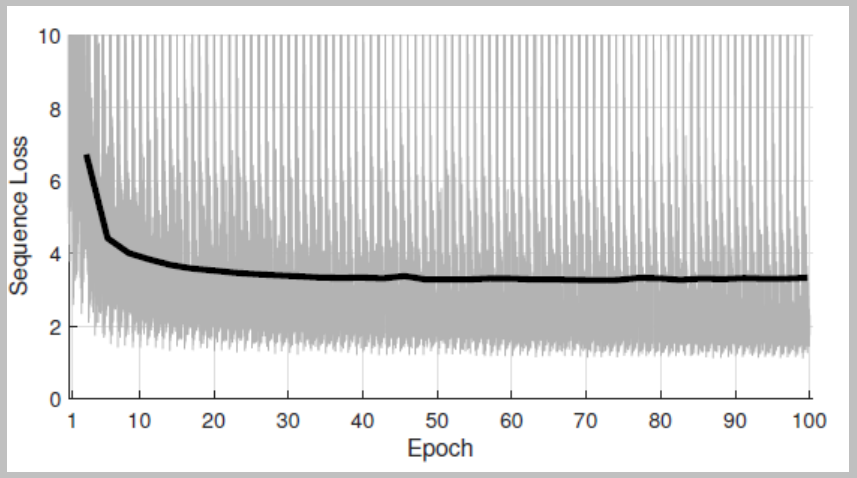
- O layer de saída (SOFTMAX) implementa a função mostrada
  - T representa uma *temperatura de amostragem*, dada pelo usuário ao rodar a rede
- No total teremos 5.599.881 parâmetros numéricos

$$\mathbf{p}_t = \text{softmax}\left(\frac{1}{T_s}\left[\mathbf{W}_s\mathbf{h}_t^{(3)} + \mathbf{b}_s\right]\right) \qquad (6)$$

# RNN-Folk - treinamento

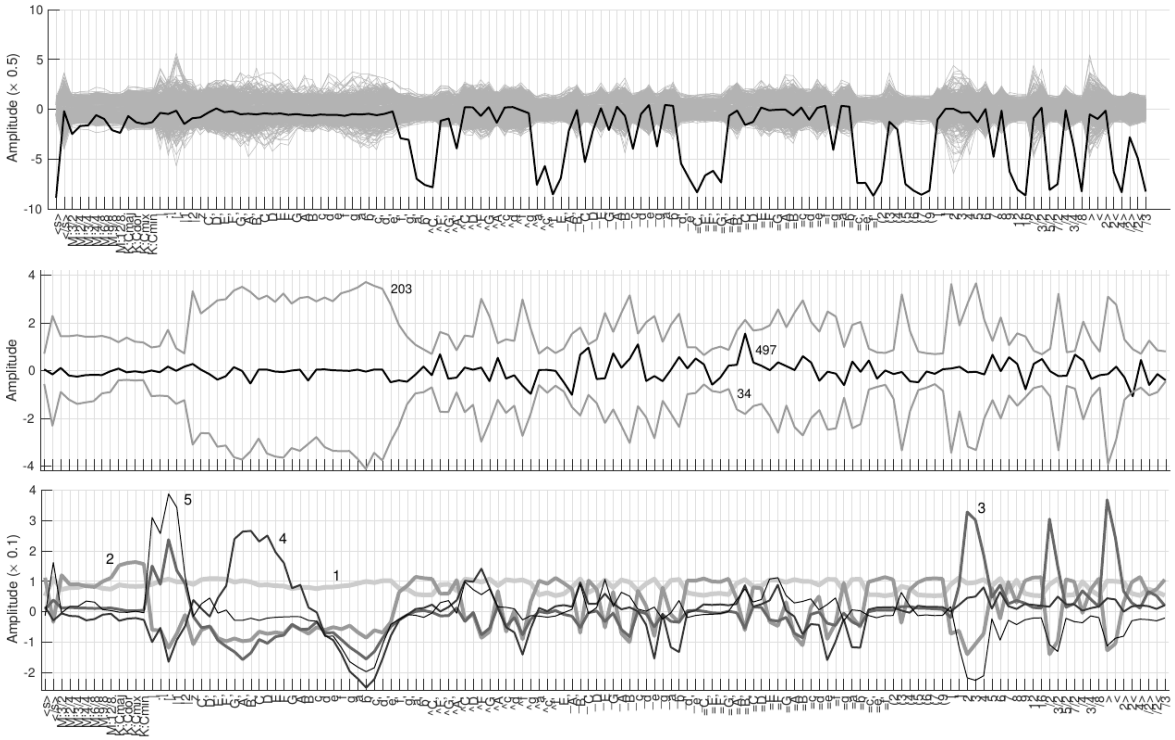
- O treinamento da rede busca minimizar a função de custo mostrada ao lado, que indica quão bem um modelo descreve o dataset.
- A figura mostra o decaimento da função de custo após 100 épocas de treinamento.

$$L(s) := -\frac{1}{|s|}\sum_{t=1}^{|s|}\log[\mathbf{p}_t]_{s(t)}$$



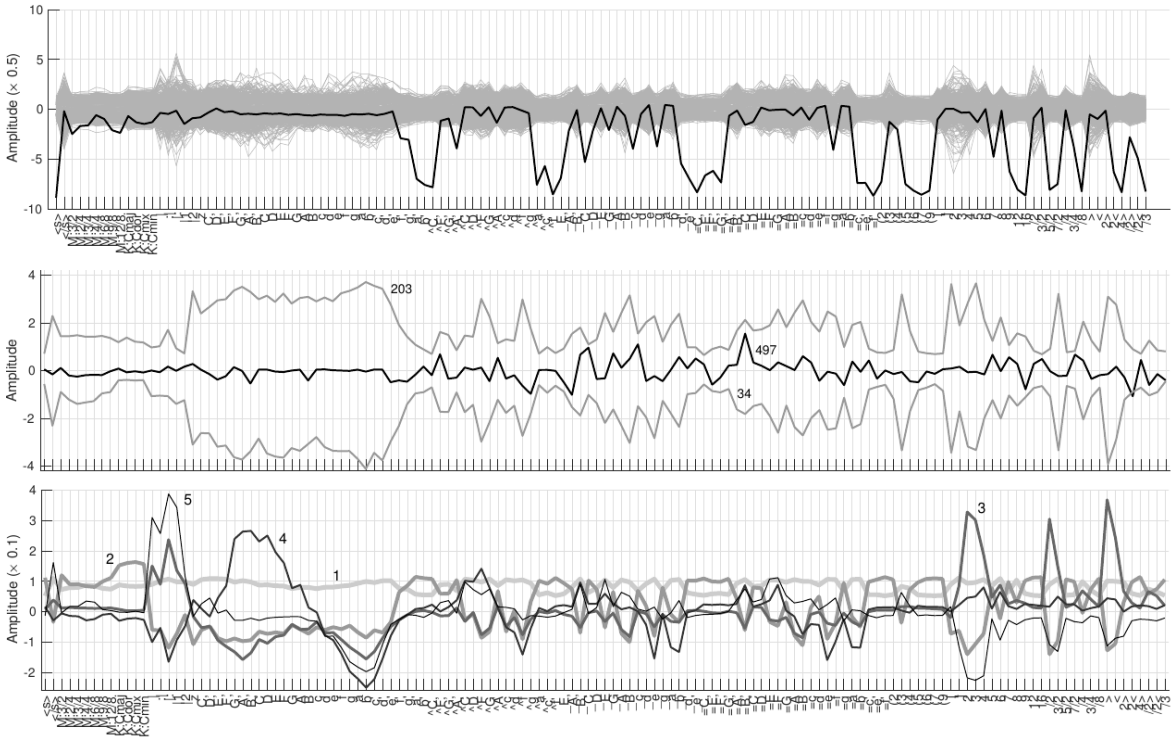
# Análise do layer SOFTMAX

- O estado oculto do 3º layer de LSTM é a entrada do SOFTMAX
  - Os gráficos mostram a amplitude de valores do LSTM contra as saídas de SOFTMAX



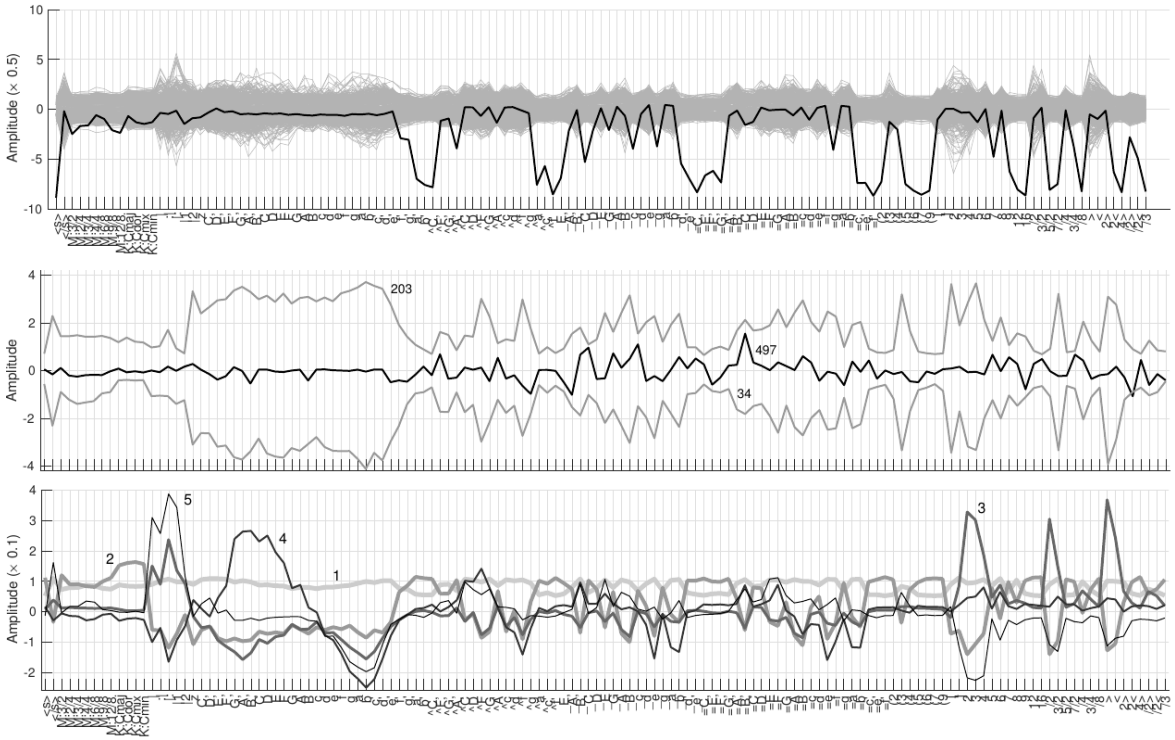
# Análise do layer SOFTMAX

- Linha 1: parâmetros $W_s$ (cinza) e $b_s$ (preto) contra o vetor de saída
  - Bias baixos relativos aos tokens <s>; =f’; 16; (7; ˆf’; =C,.
  - <s> é o token de entrada – aparece em todos os casos, possuindo semântica quase nula.
  - Os outros raramente são gerados pela rede, e quase não aparecem nos dados de treinamento.



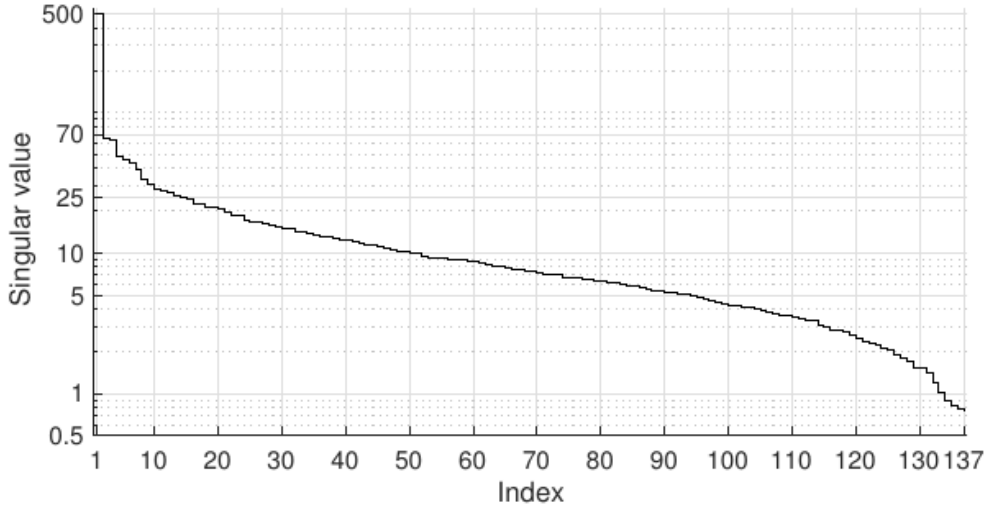
# Análise do layer SOFTMAX

- Linha 2: $W_s$ relativos às linhas 34, 203 e 497 do L3
  - Linha 497 tende a aumentar a probabilidade dos tokens =B,; _C e _c; e reduzir a probabilidade dos tokens _A e ^g (ou vice-versa)
    - Notas iguais sendo tratadas de forma similar
  - As outras linhas também apresentam probabilidades de tokens similares, mas demonstram ser opostas em polaridade
    - Diversas outras instâncias de pesos em L3 sugerem que suas colunas funcionam como grupos direcionadores do SOFTMAX
    - Baseado nisso, executou-se uma decomposição em valores singulares do vetor $W_s$



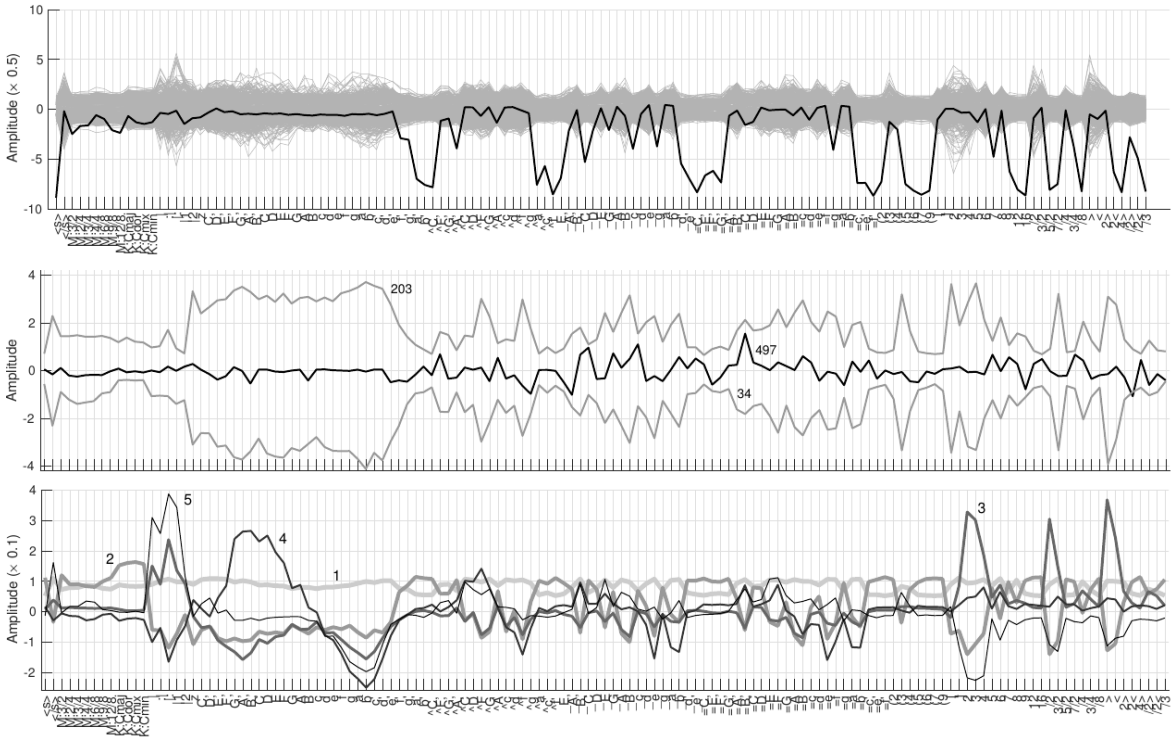
# Decomposição de $W_s$ em valores singulares

- Os maiores valores singulares respondem pela maior variância nos resultados
- Particularmente, o principal valor responde por 7,5 vezes mais variância que qualquer outro valor



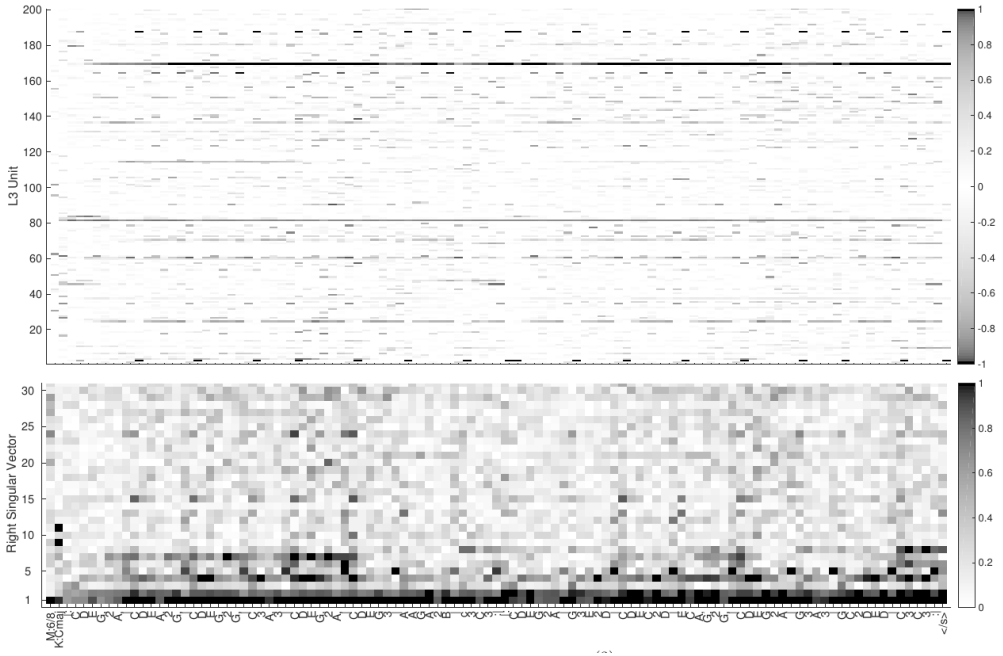
# Análise do layer SOFTMAX – valores singulares

- Linha 3: $W_S$ relativos aos 5 maiores valores singulares
  - O principal valor responde igualmente em todos os tokens, aumentando a probabilidade onde seu valor absoluto é maior e atenuando onde é menor
  - O segundo valor aumenta a probabilidade dos quatro modais e reduz dos tokens de três oitavas
  - O terceiro aumenta de diversos tokens de tempo e alguns multiplicadores (como 2, ou /2) e reduz a mesmas três oitavas do anterior
  - O quarto aumenta de alguns tokens de tons graves
  - O quinto aumenta fortemente os tokens de compasso, e atenua alguns de duração.



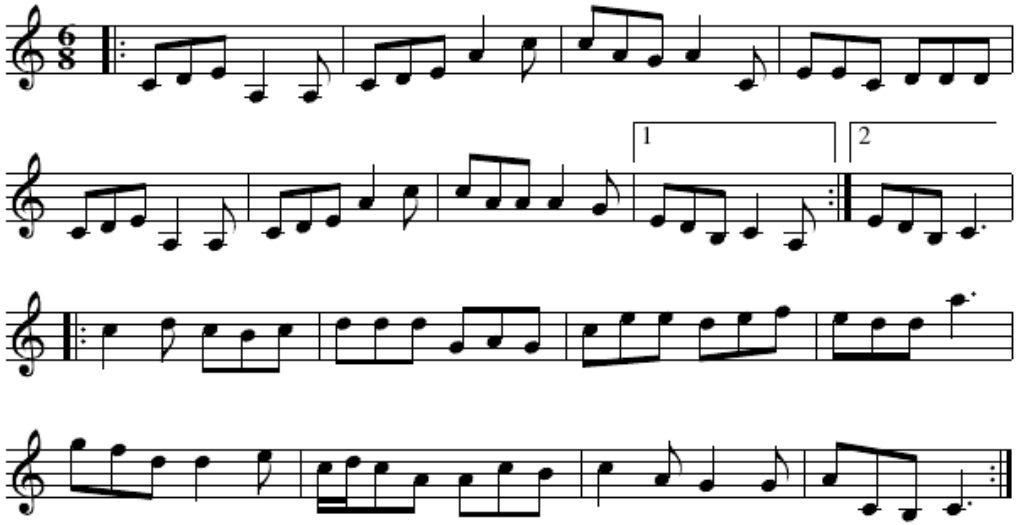
# Sequências de estados ocultos em L3

- Distribuição esparsa
- Alguns ativam na entrada de cada token de compasso, outros a cada 6 ou 7 passos (a duração mais comum de compasso), outros logo antes de um token de repetição de compasso
- Projetados sobre os 30 maiores valores singulares de $W_s$ a



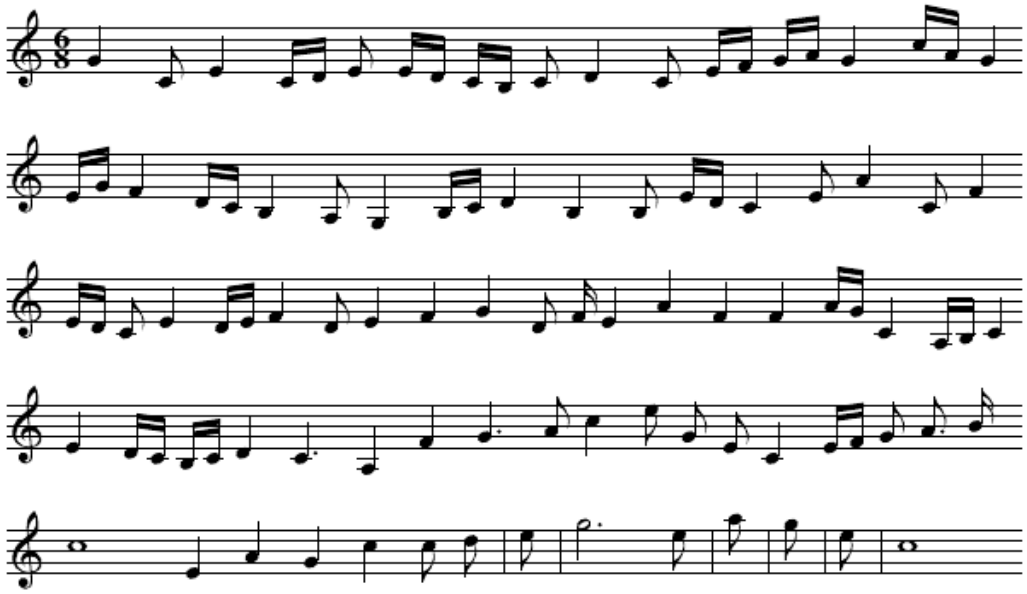
# Redução do espaço vetorial de $W_s$

- Ao truncar a soma na decomposição de valores singulares, foi reduzido o espaço vetorial para os 30 pares de vetores singulares de maior valor
- O compasso se mantém, bem como a estrutura AABB, e o início e fim em notas tônicas, mas uma maior variação no tema é introduzida

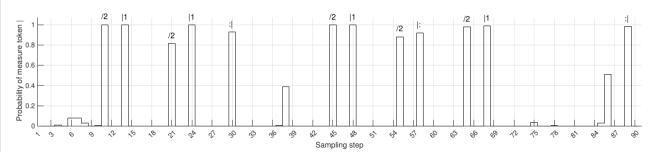


# Redução do espaço vetorial de $W_s$

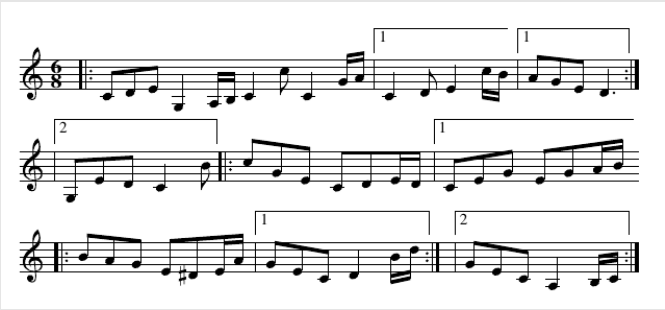
- Ao contrário, ao reduzir o espaço em 1 e alterar o SOFTMAX para Ws ← Ws − $\sigma_5 u_5 v_5^T$, a melodia perde o sentido.
- Aparentemente o componente de "direção" da melodia é severamente dependente dos tokens de compasso
- Os comportamentos internos de cada parte da melodia não parecem ser muito afetados por essa "lobotomia"



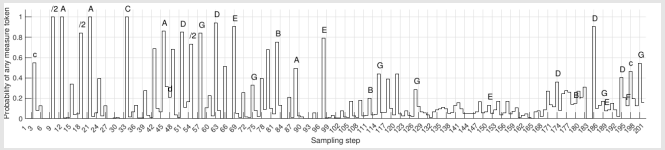
# Teste da hipótese anterior



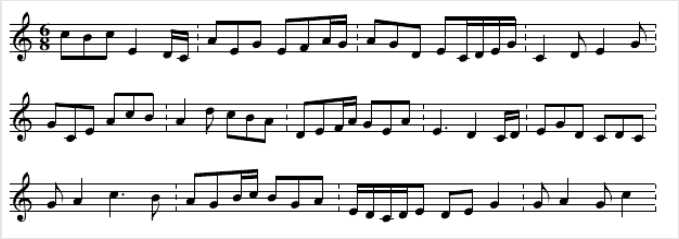

- Para testar essa hipótese, a amostragem de SOFTMAX é alterada para excluir o token |
- O resultado é uma melodia de compasso similar à original, porém sem direcionamento.



# Teste da hipótese anterior




- Ao remover todos os tokens de compasso, o resultado perde mais características, embora mantenha parte da estrutura
- O gráfico mostra que diferente do teste anterior, quase não há tentativas de encerrar o compasso.



# Deslocamento do SOFTMAX

- A partir disso, podemos generalizar o softmax de modo a criar uma expressão em que possamos avaliar a partir de cada token de saída

# Bibliografia


- Sturm, B. L. What do these 5,599,881 parameters mean?: An analysis of a specific LSTM music transcription model, starting with the 70,281 parameters of its softmax layer. *International Conference on Computational Creativity*. 2018. Disponível em http://www.diva-portal.org/smash/get/diva2:1260836/FULLTEXT01.pdf
- Sturm, B. L.; Santos, J. F.; Ben-Tal, O.; and Korshunova, I. Music transcription modelling and composition using deep learning. *1st Conference on Computer Simulation of Musical Creativity.* 2016. Disponível em https://arxiv.org/pdf/1604.08723.pdf