3
Classificação

Este capítulo apresenta primeiramente o algoritmo proposto para a classificação de áudio codificado em MPEG-1 Layer 2 em detalhes. Em seguida, são analisadas as inovações apresentadas.

3.1.
Resumo do algoritmo proposto

O algoritmo desenvolvido divide o áudio em envelopes de 4 segundos, com 2 segundos de interposição, e classifica cada envelope de acordo com as informações extraídas. Os envelopes são classificados em 4 tipos diferentes: silêncio, música, fala e aplausos. Além disso, há envelopes onde se misturam fala e aplausos. Esse fato é muito comum em programas com audiência, como programas de entrevistas, ou seriados de comédia, onde há a presença da claque. O algoritmo também detecta essas ocorrências e os envelopes são classificados como aplauso e fala. A Figura 4 ilustra a etapa inicial do algoritmo de classificação do áudio. Após essa análise, há uma etapa de pós-processamento, onde são feitos ajustes baseados em informações que requerem que a leitura e a análise de todos os envelopes já tenham sido realizadas. Nos próximos itens vamos analisar cada etapa do algoritmo detalhadamente.

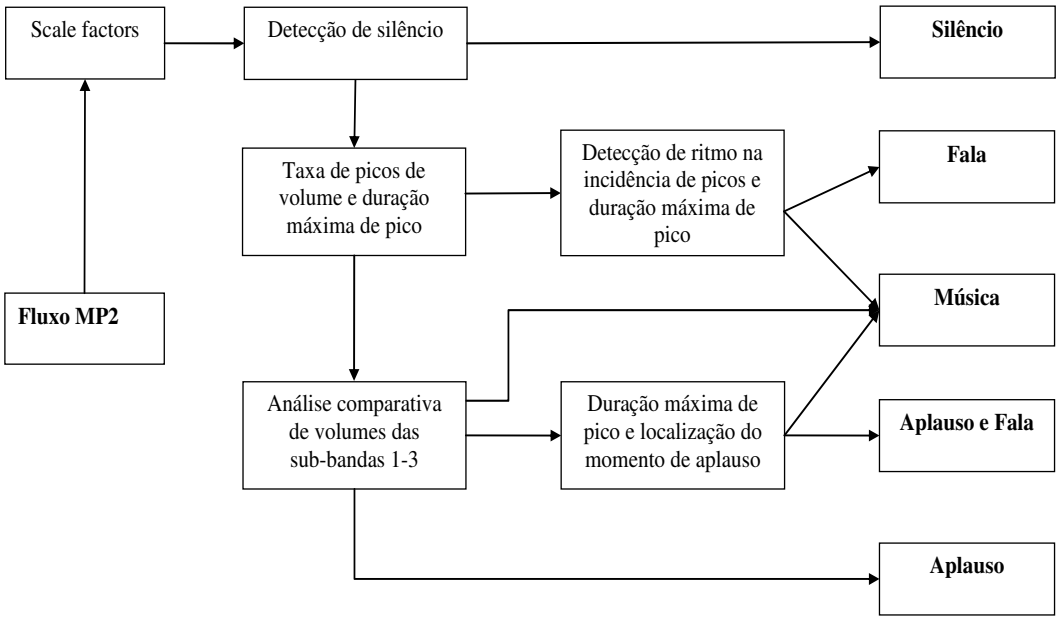

Figura 4 – Algoritmo de classificação do áudio

3.2.
Características do áudio extraídas para análise

É muito comum haver pequenos ruídos no áudio, imperceptíveis ao ouvido humano, por serem de muito curta duração. Para evitar que esses ruídos causem distorções na análise do áudio, todo o processamento dos valores dos *scale factors* é feito a partir de uma média do *frame* corrente com seus 2 *frames* adjacentes em cada direção.

Para cada envelope, são extraídas as seguintes características:

- Picos por segundo: picos são definidos como intervalos onde o volume médio do áudio é maior que um determinado limite. Após os testes, o limite que teve os melhores resultados foi o de metade da média do envelope. Logo, cada bloco de áudio onde todos os *frames* têm volume acima desse limite é considerado um pico. Para essa análise, são utilizados apenas os valores das sub-bandas de 2 a 7 do áudio.

- Pico mais longo: são calculadas as durações de cada pico encontrado no envelope e é armazenado o valor do pico mais longo. Para essa análise, são utilizados apenas os valores das sub-bandas de 2 a 7 do áudio.

- Sub-bandas 1-3: são extraídas as médias dos *scale factors* das sub-bandas de 1 a 3. Armazenam-se as médias gerais e médias locais para cada bloco de 0,5 segundos, totalizando 8 blocos por envelope.

- Sub-bandas 14-16: são extraídas as médias dos *scale factors* das sub-bandas de 14 a 16. Armazenam-se apenas as médias gerais de cada envelope.

- Periodicidade de picos: a partir dos picos detectados, é feita uma análise calculando o grau de periodicidade da incidência de picos. Para isso, inicialmente é calculada a média das distâncias entre picos. Essas distâncias são calculadas a partir dos intervalos entre o fim de um pico e o início do pico seguinte. Considerando o valor dessa média, é calculada a soma da distância dos picos do envelope em relação à média, somando-se o quanto cada intervalo se distancia dela. O resultado da soma é considerado o grau de periodicidade de picos do envelope, que é medido em milissegundos.

É também calculado o nível de silêncio, que é comum a todos os envelopes. Calcula-se esse nível a partir da média geral de volume dos envelopes, avaliando-o em 7,3% desse valor.

3.3. Funcionamento do algoritmo

A partir das informações extraídas do áudio descritas na seção anterior, para cada envelope é feito um processamento inicial, que resulta em uma classificação. Em seguida, ocorre o pós-processamento, onde ajustes são realizados nessa classificação.

Além das classificações que são apresentadas como resultado final da análise, o algoritmo trabalha com uma classificação a mais, interna, chamada *beat music*. Essa classificação foi desenvolvida devido a que, muitas vezes, músicas com batidas/tempo marcante, como música eletrônica, têm características que se assemelham muito às da fala. Assim, o algoritmo pode considerar internamente a maneira como esse envelope foi classificado como música, e considerar isso na etapa de pós-processamento, onde pequenas distorções na análise podem ser corrigidas.

3.3.1. Etapa principal

Nesta etapa ocorre a extração das informações listadas no item 3.2. Além disso, é feito o procedimento inicial de classificação para todos os envelopes, detalhado a seguir.

O primeiro passo do algoritmo é a detecção de silêncio. Para isso, compara-se a média do volume do envelope com o nível de silêncio, calculado anteriormente. Caso o volume do envelope seja menor que esse nível, esse é considerado como silêncio e o algoritmo passa para o próximo envelope sem necessidade de realizar o restante dos passos.

Após a detecção de silêncio, é feita uma análise comparativa dos valores das sub-bandas 14 a 16 com a sub-banda 1. As freqüências representadas por essas sub-bandas em geral têm volume comparativo mais alto em música (principalmente música clássica) do que em outros tipos de áudio. Compara-se o valor das médias de cada uma dessas sub-bandas com o valor da sub-banda 1 dividido por 20. Se para duas ou mais dessas sub-bandas o valor for maior do que

o da sub-banda 1 dividido por 20, é setado um *flag* indicando essa ocorrência. Esse *flag* é utilizado no pós-processamento.

Em seguida, são analisados os valores de pico mais longo e quantidade de picos por segundo. Caso o pico mais longo seja menor do que 1 segundo e o valor de picos por segundo seja maior ou igual a 1,0 e menor que 5,5, temos indicativo de fala. Nesse caso, é feita a análise do valor de periodicidade, para classificar o envelope entre fala e *beat music*, e o algoritmo passa para a análise do próximo envelope. Em caso contrário, a classificação ainda está indefinida e são realizadas as próximas etapas. Para que o áudio seja considerado *beat music*, o valor de periodicidade calculado deve ser menor ou igual a 75 e o pico mais longo deve ter duração superior a 0,55 segundos. Valores pequenos na periodicidade indicam um ritmo bem definido que é característico de música, principalmente no caso de música eletrônica, que é orientada pela batida. Se isso não ocorrer, o envelope é classificado como fala.

Na próxima etapa, o áudio é avaliado procurando a incidência de aplausos. Para isso, são analisados os valores calculados para as sub-bandas 1 a 3, que em caso de aplausos possuem uma concentração maior nas sub-bandas 2 e 3 do que o observado em música e fala. Caso a média geral de volume da sub-banda 2 seja maior que a da sub-banda 1, o envelope é classificado como aplauso e o algoritmo passa para a análise do próximo envelope. Outro caso classificado como aplauso é se tanto o volume geral da sub-banda 2 quanto o da sub-banda 3 são maiores que a metade do volume geral da sub-banda 1. Se nenhuma dessas hipóteses ocorrer, o algoritmo passa para a análise individual dos blocos de 0,5 segundos dessas sub-bandas.

Nessa etapa, são feitos testes semelhantes aos da etapa anterior, porém são analisados os blocos correspondentes, ao invés do envelope inteiro. Para cada bloco, é testado se o volume médio da sub-banda 2 é maior do que o da sub-banda 1. Se isso é verdadeiro, o bloco é considerado como possível bloco de aplausos. O mesmo ocorre caso os volumes médios do bloco nas sub-bandas 2 e 3, ambos multiplicados por 1,5, sejam maiores que o volume médio do bloco na sub-banda 1. Caso sejam encontrados pelo menos dois blocos consecutivos com indicação de aplausos e o pico mais longo do envelope seja menor do que 2,5 segundos, considera-se que parte do envelope pode ser de aplausos, e a análise do envelope passa para o próximo passo. Nesse caso, é também setado um flag indicando que

o envelope contém indicação de aplausos, a ser utilizada na etapa de pós-processamento do algoritmo. Em caso contrário, ou não há indicação suficiente de aplausos ou a duração do pico mais longo é típica de música. O envelope é então considerado como de música e o algoritmo passa para a análise do próximo envelope.

Em seguida, é avaliado se o intervalo de palmas encontrado coincide com o pico mais longo do envelope. Caso esse intervalo comece ou termine dentro do pico mais longo, é considerado que há coincidência, e continua-se a análise do envelope. Caso contrário, o envelope é classificado como música e passa-se para o próximo envelope.

No passo final dessa etapa da classificação, avaliam-se as alterações na taxa de picos do envelope durante o intervalo onde foi encontrada possível incidência de aplausos. Caso seja confirmada a ocorrência de aplausos como um fator de alteração significativa da taxa de picos, sabemos que o restante do envelope é composto de fala, e então ele será classificado como fala e aplausos. Para tanto, são calculados os valores totais de picos durante o intervalo onde há indicação de aplausos e fora desse intervalo. Se a quantidade de picos durante o intervalo de aplausos for menor ou igual a 1 e a quantidade de picos fora desse intervalo for maior ou igual a 4, o envelope recebe a classificação de fala e aplausos e o algoritmo passa para o próximo envelope. Caso contrário, é calculada a taxa de picos por segundo para cada um desses intervalos. Se a diferença dessas taxas for maior do que 1, o envelope é classificado como fala e aplausos. Caso isso não ocorra, o envelope é classificado como música.

3.3.2. Pós-processamento

Nesta etapa são feitos ajustes na classificação dada aos envelopes na etapa anterior. Com as informações já obtidas na primeira etapa e o conhecimento da sobreposição que ocorre entre os envelopes, podem-se avaliar melhor os dados relativos a alguns envelopes, que podem então ter sua classificação alterada. Um exemplo é o caso de um envelope de música entre vários de fala e vice-versa. Por causa da sobreposição dos envelopes, essas seqüências de envelopes são muito raras de acontecer, e tais ocorrências são tratadas no pós-processamento. Outros

casos são o *flag* indicando música, que foi preenchido durante a primeira etapa do processo, e os envelopes com a classificação beat music ou com o *flag* setado indicando a presença de aplausos em parte do envelope. Como essa classificação pode ter um grau de incerteza, em alguns casos ela requer uma análise dos envelopes adjacentes, o que é feito no pós-processamento. Todos esses procedimentos serão detalhados abaixo, na descrição do algoritmo.

Inicialmente, é avaliada a incidência do *flag* que indica música. Caso ele esteja habilitado em 40% ou mais dos envelopes do vídeo, um novo *flag* é habilitado, indicando que deve haver troca da classificação para música dos envelopes marcados com esse *flag* no processamento de cada envelope. Em seguida, são aplicados os passos abaixo a cada envelope.

O primeiro passo é baseado no *flag* que indica música. Caso o *flag* que indica que deve haver troca esteja marcado e o *flag* que indica música no envelope esteja habilitado, a classificação do envelope é alterada para música, e o algoritmo passa para o próximo envelope.

Em seguida, é avaliado se o envelope tem algum grau de incerteza na sua classificação. Tanto envelopes classificados como beat music quanto envelopes onde o *flag* indicando a presença de aplausos em parte do envelope é verdadeiro são considerados com a classificação apresentando incerteza. Esses envelopes recebem no próximo passo uma análise diferente da aplicada aos envelopes para os quais isso não ocorre.

Para os envelopes com incerteza, é feita uma análise dos envelopes adjacentes, englobando os dois envelopes anteriores e os dois seguintes ao que está sendo avaliado. Desses envelopes, é calculado o total classificado como música (música ou beat music), o total classificado como aplausos ou fala (fala, aplausos ou fala e aplausos) e um valor que é definido como não-processados, que são os envelopes que ainda não foram analisados pelo pós-processamento (das adjacências seguintes) e cujo *flag* que indica a presença de aplausos em parte do envelope está setado. Caso o total classificado como aplausos ou fala seja maior ou igual a 3 e, ou o total classificado como música seja 0 ou o total de não-processados seja 1, a classificação do envelope é alterada para fala e aplausos. A mesma avaliação é feita para o caso de música. Caso o total de envelopes classificado como música seja maior ou igual a 3 e, ou o total classificado como

aplausos ou fala seja 0 ou o total de não-processados seja 1, a classificação do envelope é alterada para música.

Para os envelopes que não possuem incerteza a análise é diferente: é testado se um envelope de música está entre dois de fala, ou se um envelope de fala está entre dois de música. Por causa da interposição, isso é muito difícil de acontecer na realidade. Quando acontece, a classificação do envelope corrente é substituída pela dos dois envelopes adjacentes. Um caso especial é quando ocorre uma seqüência de envelopes trocando entre música e fala. Caso essa seqüência seja maior do que 5 envelopes, o algoritmo considera que isso é uma característica do vídeo que está sendo analisado, e não faz a troca enquanto essa seqüência durar.

3.4. Inovações apresentadas

O algoritmo de classificação baseia-se em uma adaptação do método apresentado nos artigos do grupo da Universidade de Dublin, descritos no capítulo anterior. As idéias desses trabalhos foram usadas como base para o algoritmo proposto e adaptadas e ajustadas em busca de melhores resultados. Foram também desenvolvidas novas técnicas para melhorar a eficiência do algoritmo e incluir a detecção de aplausos, sem que a precisão fosse afetada pela inclusão da nova categoria.

A detecção de silêncio é inteiramente baseada na técnica proposta nesses artigos. Ela provou ser eficiente, dispensando alterações. Fora isso, foram extraídas as idéias básicas dos artigos, mas elas sofreram alterações na maioria dos casos.

Conforme explicado anteriormente, o artigo sugere o uso de envelopes de 4 segundos, com 2 segundos de interposição. Foram também usadas as idéias de análise da taxa de picos e do pico mais longo dos envelopes, baseados apenas nas sub-bandas de 2 a 7. Os parâmetros utilizados no artigo foram alterados a partir de testes, buscando maior precisão na análise e de modo a se ajustarem às novas técnicas desenvolvidas.

A primeira nova técnica desenvolvida foi o cálculo da periodicidade dos picos. Essa técnica baseia-se na prerrogativa de que músicas, principalmente eletrônicas, respeitam um tempo pré-definido. Em músicas desse estilo, onde a batida é muito forte, é comum que as batidas que marcam o tempo sejam detectadas como picos curtos. Isso induzia o algoritmo ao erro, pois tais músicas apresentavam características que eram confundidas com as da fala. A análise de periodicidade detecta justamente essas ocorrências e permitiu corrigir a classificação de músicas com batida forte, enquanto anteriormente erros eram muito comuns.

Outra inovação importante foi a detecção de aplausos. A adição de aplausos como uma classificação é importante não só para essa tarefa, como adiciona muito à segmentação. Aplausos servem como indicativos de eventos importantes em diferentes tipos de vídeo. Em um concerto ou apresentação musical, por exemplo,

marcam o começo da apresentação, ou o fim da performance de uma música. Em programas de entrevistas, indicam apresentações de convidados, fins de entrevistas, ou intervenções do público que podem apontar momentos que devem ser indexados. Já em seriados de comédia para televisão, o uso da claque é muito comum e indica piadas ou mudanças de cena que ocorrem após as risadas da claque.

Há diversos métodos na literatura para detectar aplausos, porém nenhum deles trabalha apenas com os *scale factors* dos *frames* de áudio. Com a informação reduzida por esse fato, foi feita uma análise detalhada das alterações no áudio em diversos tipos de vídeo com aplausos. A partir dessa análise foi desenvolvido o método de detecção de aplausos, que se destaca por sua simplicidade. Quase nenhum processamento adicional é envolvido, pois os valores dos *scale factors* das sub-bandas já são lidos quando os envelopes de áudio são montados. Esse método detecta aplausos com grande precisão, e permite não só a detecção de envelopes compostos inteiramente de aplausos como de envelopes que são parcialmente aplausos e parcialmente fala, que têm muitas características semelhantes a envelopes de música. A análise bloco a bloco e os ajustes realizados no pós-processamento evitam que esses erros ocorram.

A utilização das sub-bandas 14 a 16 como indicativo de música é outro fator importante. Essa técnica parte de um princípio similar ao da técnica de detecção de aplausos, por trabalhar diretamente a partir dos *scale factors*. Ela foi desenvolvida para aumentar a eficiência, especialmente da análise de áudios de música clássica, que muitas vezes se mostrava ineficaz para músicas longas, como apresentações de orquestras, que têm muitas variações, e por vezes possuem características que levavam a uma classificação errônea. Essa técnica se baseia no fato de que a música em geral atinge freqüências e, conseqüentemente, sub-bandas mais altas do que a fala, com um nível significativo. Com ela, a taxa de acertos em vídeos onde o tipo de áudio é música clássica aumentou de 67% para 87%.

A última inovação apresentada se encontra no pós-processamento. A maioria das técnicas analisadas não se utiliza das informações dos *frames* adjacentes, que são fortes indícios do que ocorre em um frame numa técnica com sobreposição de envelopes como a apresentada nesta dissertação. Assim, o algoritmo desenvolvido para o pós-processamento se utiliza dessas informações para auxiliar a tomada de decisões. Em envelopes onde a classificação apresenta

algum grau de incerteza os erros de classificação são muito comuns, então o uso dessa técnica diminui a quantidade de erros causados por pequenos desvios no padrão do áudio. O pós-processamento aumenta a taxa de acertos do algoritmo em até 3% para áudio diferente de música clássica.