2
Trabalhos relacionados

Há diversas abordagens na literatura para os problemas analisados nesta dissertação. Podemos dividir os trabalhos pesquisados inicialmente entre os que trabalham no domínio comprimido e os que fazem a reconstrução da onda sonora e trabalham a partir de informações retiradas dessa onda. Os últimos trabalhos em geral não consideram o padrão de áudio que está sendo lido, e serão analisados na seção 2.1. No domínio comprimido, há estudos a partir de diferentes padrões. A seção 2.2 trata dos estudos em domínio comprimido, porém que trabalham em domínios diferentes dos fatores de escala do MPEG-1 Layer 2. Esses algoritmos fazem a leitura de mais informações do que a é realizada nesta dissertação, e têm resultados interessantes. Já a seção 2.3 apresenta os trabalhos encontrados que usam apenas as informações dos *scale factors* e mais se assemelham a esta dissertação.

2.1.
Domínio não-comprimido

Nesse domínio é possível a obtenção de todas as informações do áudio, inclusive a onda sonora reconstruída. Sendo assim, os trabalhos aqui situados abdicam de maior eficiência em termos de tempo para obter altos graus de precisão em suas análises. Uma vantagem deles é que, por trabalharem a partir da onda sonora, o padrão de codificação do áudio não é importante.

R.-S. Lin e L.-H. Chen (2005) propõem a classificação do áudio a partir de sua decomposição usando *wavelets* de Gabor. O áudio é transformado em um espectrograma, do qual são extraídos diferentes padrões através de filtros e *wavelets* de Gabor. Em seu trabalho, eles atingem uma precisão de 98% para distinção entre fala e música, e 95% para a classificação em cinco tipos (fala pura,

música instrumental, música com voz, fala com música de fundo e fala com ruídos no fundo).

J. Foote (2000) apresenta um método para a segmentação a partir da análise no domínio de freqüências da onda sonora. Nesse trabalho, ele se baseia na avaliação da auto-similaridade local do fluxo de áudio para procurar mudanças de cena. Assim, analisa o quanto um instante do áudio é semelhante aos instantes mais próximos dele. Esse método trabalha diretamente a partir da onda sonora, que é transformada para o domínio de freqüências a partir de uma transformada rápida de Fourier. São analisadas a auto-similaridade no passado e futuro e a similaridade entre o ponto passado e o ponto futuro. Quando a auto-similaridade é alta e a similaridade entre o passado e o futuro é baixa, o algoritmo considera tal ponto como um segmento.

O artigo ainda ilustra a possibilidade do uso de diferentes janelas para a definição do ponto passado e do ponto futuro. Esse tamanho de janela determina o nível de detalhes no qual o algoritmo deve trabalhar – desde notas musicais para uma janela muito pequena até passagens musicais longas para janelas grandes.

O método de Foote serviu de inspiração para o algoritmo de segmentação proposto nesta dissertação, por sugerir a análise do áudio comparando a similaridade de momentos.

Y. Zhu e D. Zhou (2003) sugerem um método que faz a classificação do áudio a partir de diversas características extraídas da onda sonora e combina essas informações com a análise da imagem para conseguir uma precisão acima de 90% na segmentação do vídeo.

A análise do áudio primeiramente o classifica em quatro tipos: silêncio, som ambiente, música e fala. Para isso, o método extrai quatro tipos de características do áudio: média de energia em tempo curto, taxa de cruzamento por zero em tempo curto, freqüência fundamental e centróide de freqüência. O silêncio é caracterizado por baixa energia em tempo curto e baixa taxa de cruzamento por zero. O centróide de freqüência é um bom indicativo para sons ambientes, pelo fato de muitos sons desse tipo terem maior concentração em freqüências altas em comparação com música e fala. A taxa de cruzamento por zero é usada para distinção entre música e fala, por ter diferenças significativas devido às características dos sinais de fala, com muitos picos curtos, e música, com poucos picos longos.

Além disso, o método se utiliza de um Critério de Informação Bayesiano para localizar mudanças de interlocutor. Segmentos do áudio são divididos a partir da mudança de classificação do áudio, ou da detecção de mudança de interlocutor.

Os autores descrevem também como é feita a segmentação a partir do vídeo. As mudanças de cena encontradas por áudio e vídeo são combinadas para se extrair o resultado final do algoritmo. Num primeiro passo, quando ambas as fontes coincidem, o instante é considerado uma possível mudança de cena. Quando o segmento do áudio é originado por uma mudança de classificação, é confirmada a mudança de cena. Quando a mudança de cena é originada de uma mudança de interlocutor, é feita uma análise da seqüência de imagens próxima à fronteira da cena. Se os objetos presentes na imagem ou o ambiente de fundo da imagem têm um grau de correlação alto, o algoritmo considera a possibilidade de que esteja ocorrendo um diálogo, por exemplo, e assim esse momento não é considerado uma mudança de cena. Caso contrário, a fronteira encontrada é confirmada.

2.2. Domínio comprimido

A maioria dos métodos que trabalham no domínio comprimido utiliza-se dos valores das amostras do áudio, porém sem fazer a reconstrução da onda sonora. Assim, há um ganho em desempenho em relação aos métodos analisados na seção 2.1, porém as características que podem ser extraídas do áudio são mais limitadas.

Y. Nakajima, Y. Lu, M. Sugano, A. Yoneyama, H. Yanagihara e A. Kurematsu (1999) propõem um método de classificação a partir dos valores reconstruídos das sub-bandas de um fluxo de áudio MPEG. Esse método classifica o áudio em música, fala ou aplausos, e atinge 90% de acerto para detecção de música e fala. Para isso, ele se utiliza de informações sobre a densidade temporal, a largura de banda e a freqüência central da energia das sub-bandas para fazer sua classificação. O áudio é dividido em blocos de 1 segundo, classificados independentemente.

Para a detecção de silêncio, os autores sugerem o uso da variância da energia na sub-banda 0. Foi observado no trabalho que essa variância é

significativamente menor em intervalos de silêncio. A distinção entre música e fala é baseada na densidade de energia temporal e na largura de banda. Os autores observam que a distribuição de energia na fala é intermitente, devido às pausas que ocorrem, enquanto na música ela é mais uniforme. Já na música, a largura de banda é normalmente maior do que na fala. A quantidade de sub-bandas com um volume significativo na música é consistentemente maior do que na fala. Assim, analisam-se detalhadamente os níveis de energia em cada *frame* do intervalo analisado para observar essas características e fazer a distinção entre música e fala.

Aplausos são fortes indicativos de começo e fim de músicas em concertos, entrevistas ou apresentações em *talk-shows*, e cenas ou episódios em seriados de televisão. Suas características são únicas entre as possíveis classificações. Quando comparado com música ou fala, os aplausos têm uma similaridade contínua, ao contrário das muitas variações que os outros dois apresentam, e a freqüência central é mais estável. Assim, o método proposto nesse artigo calcula o centróide de sub-banda de cada frame, para poder estabelecer sua média e variação para análise.

Para finalizar a classificação, os autores se utilizam de uma função discriminante de Bayes para distribuição Gaussiana multivariada.

S. Venugopal, K.R. Ramakrishnan, S.H. Srinivas e N. Balakrishnan (1999) usam a discriminação entre fala e música como base para a segmentação, para a qual também se utilizam de um método para diferenciar interlocutores e seus gêneros. Os autores determinam três mudanças de características do áudio como indicativos de mudanças de segmentos. São estes: transição do áudio entre música e fala, mudança de interlocutor e mudança do gênero do interlocutor.

Para a classificação do áudio, o artigo propõe que sejam extraídas diversas características do fluxo de áudio, e a partir dos indicativos de cada uma é avaliada a probabilidade do áudio ser música ou fala, e assim definida a classificação. São elas: tonalidade, largura de banda, padrões de excitação, duração tonal e seqüências de energia.

A tonalidade se baseia no fato de que a música é geralmente composta de uma multiplicidade de tons, com distribuições únicas de harmônicos, enquanto a fala exibe seqüências alternantes de segmentos tonais e de ruídos. A utilização da largura de banda é a mesma observada no artigo de Nakajima et al. (1999), onde

se observa que o áudio possui uma largura de banda maior que a fala. Os padrões de excitação são definidos pelo fato da fala normalmente estender-se por apenas três oitavas, enquanto uma música normalmente estende-se por aproximadamente seis oitavas. Uma oitava é o intervalo entre uma nota musical e outra com a metade ou o dobro de sua freqüência. Na fala, pode ser observada uma taxa silábica, pelo fato da ocorrência de vogais ser regular. Isso é considerado na duração tonal. Já as seqüências de energia partem da observação que a fala segue um padrão no qual segmentos de alta energia são seguidos por segmentos de baixa energia, o que não é comum na música.

Para a identificação de interlocutores, o artigo sugere a utilização de um método a partir de um modelo de mistura Gaussiano. Foi criado um modelo para cada interlocutor de interesse nos testes realizados.

A identificação de gênero, por sua vez, é feita a partir de uma combinação de dois métodos. Um deles utiliza-se novamente de um modelo de mistura Gaussiano, treinado a partir de locutores de cada sexo. O outro utiliza-se de uma técnica para avaliar o timbre do interlocutor. Como o timbre masculino está em uma faixa de freqüências diferente do feminino, a identificação da freqüência onde o timbre atua permite a identificação do gênero do interlocutor.

Esses métodos atingem uma precisão de aproximadamente 75% para identificação de interlocutores a partir de seus modelos, e 88% para classificação correta de segmentos de fala.

S. Kiranyaz, M. Aubazac e M. Gabbouj (2003) fazem a classificação em fala, música ou silêncio e segmentação a partir desses fragmentos do áudio nos padrões MP3 ou AAC. O método sugerido funciona para os dois padrões por ambos trabalharem uma transformada de cossenos (MDCT), e as informações serem retiradas a partir dos valores resultantes dessa transformada.

As características do áudio analisadas são: energia total do *frame*, taxa de energia por banda, freqüência fundamental, centróide de sub-banda e taxa de pausas. O algoritmo proposto primeiramente classifica os *frames* individualmente, para obter uma segmentação inicial a partir dos segmentos considerados silêncio e não-silêncio. Então, são usadas a taxa de energia por banda e a taxa de pausas para avaliar os segmentos de não-silêncio entre música e fala. Em seguida, segmentos de silêncio muito curtos são eliminados e incorporados aos segmentos adjacentes, que são reclassificados. Quando todos os segmentos de silêncio considerados

pequenos demais são eliminados, é feita uma classificação final dos segmentos não-silêncio encontrados, a partir da freqüência fundamental, do centróide de sub-banda e da taxa de pausas. Em seguida, os segmentos não-silêncio são novamente analisados em busca de uma segmentação adicional, que não parta de um intervalo de silêncio para dividir o segmento entre música e fala. Os segmentos são divididos ao meio, buscando-se uma diferença significativa no centróide de sub-banda. Caso seja encontrada, são realizadas subdivisões adicionais até que se encontre a fronteira correta entre os segmentos, que são então re-classificados.

2.3. Trabalhos a partir dos *scale factors*

Há muito poucos trabalhos disponíveis na literatura que trabalham nos *scale factors* do Layer 2 do MPEG. Os únicos trabalhos significativos encontrado foram do grupo da Universidade de Dublin, explicados em Sadlier et al. (2001), Jarina et al. (2001), Jarina et al. (2002a), Jarina et al. (2002b) e Jarina et al. (2004). Esses trabalhos sugerem a classificação do áudio a partir dos valores de pico mais longo e quantidade de picos para intervalos de 4 segundos, dividindo o áudio entre silêncio, música e fala com uma precisão que varia de 86% para música instrumental até 98% para fala sem ruídos de fundo.

Em Sadlier et al. (2001), é sugerido um método para a detecção de silêncio. Esse método baseia-se no cálculo do volume médio do áudio durante todo o vídeo. O limite de silêncio é calculado a partir de uma porcentagem desse volume médio, estabelecida nesse trabalho em 7,3%. Assim, *frames* onde o volume é menor do que esse limite são considerados *frames* de silêncio.

O método para a classificação é detalhado em Jarina et al. (2001). O áudio é dividido em envelopes de 4 segundos, com 2 segundos de interposição entre os envelopes. São extraídas do áudio duas características que vão servir como base para a avaliação: a taxa de picos por segundo e a duração do pico mais longo de cada envelope. A partir dessas duas informações, o processo sugerido é bem simples. Caso a duração do pico mais longo seja menor do que 0,7 segundos e a taxa de picos por envelope esteja entre 2,5 e 5,5 picos por segundo, o envelope é considerado fala. Em qualquer outro caso, é considerado música. São feitos

experimentos com diferentes limites para o estabelecimento de picos, e o que tem melhor resultado para os vídeos estudados é o de 40% do volume médio.

O algoritmo também sugere uma etapa de pós-processamento, que corrige erros singulares de classificação. Nela, envelopes de música que se encontram entre dois envelopes de fala, e vice-versa, são alterados para receber a classificação dos envelopes adjacentes, pelo fato de tal ocorrência ser extremamente rara.

Esses trabalhos serviram de inspiração para o algoritmo de classificação proposto no capítulo 3 desta dissertação, que baseou-se nos resultados encontrados por esse grupo para seu desenvolvimento inicial, utilizando-se também das características de pico mais longo e quantidade de picos por segundo para os blocos de classificação.